Statistical Methods for Air Quality Model Calibration and Validation in Urban Areas

Lauren Holmes Sim

Submitted in fulfilment of the requirements for the Degree of Master of Science

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

2016

Abstract

It is thought that air quality modelling is vital, this is because of the lack of monitoring sites and diffusion tubes in cities making it difficult to see spatially how air pollution is behaving. In this thesis, the Atmospheric Dispersion Modelling System (ADMS)-Urban is focussed on and it is of interest to see how well the modelled nitrogen dioxide (NO_2) predictions and the monitoring site NO_2 data are calibrated in Aberdeen over the year 2012. There are only six monitoring stations in Aberdeen and it will be highlighted in this thesis how close in space these monitoring stations are. To evaluate how comparable the modelled and monitoring data are, methods such as Deming Regression, Extreme Value Analysis, Functional Principal Components Analysis (FPCA) and Clustering and Functional Regression will be investigated. FPCA and clustering and Deming Regression highlight that the modelled and monitoring data appear not very well calibrated at Wellington Road, however these data are reasonably well calibrated at the other monitoring sites. FPCA and Clustering indicate that the roads appear to dominate and be the main cause of concern in Aberdeen in terms of air pollution. All methods suggest that between April 9^{th} and July 18^{th} the model and monitoring data appear not be well calibrated and this could be further explored to examine the potential causes. These analyses have identified how the relationships between the ADMS-Urban model output and the observed data may vary over time and space.

Contents

1 Introduction **1**

1.1 Air Pollution and the Effects on Human Health 1

1.2 Current Air Pollution Guidelines 2

1.3 The City of Aberdeen 5

1.4 Air Quality Modelling 6

1.4.1 Air Modelling for DEFRA 6

1.4.2 Atmospheric Dispersion Modelling System 8

1.5 Description of the data 10

1.6 Aims 14

2 Model Measurement Comparisons **16**

2.1 Model Description 18

2.1.1 ADMS-Urban Model 18

2.2 Methodology 22

2.2.1 EIV Regression 23

2.2.2 Bland Altman Plots 25

2.2.3 Extreme Value Theory 25

2.3 Comparing Measured Data and ADMS-Urban Modelled Data 28

2.3.1 Monthly timescale 30

2.3.2 Daily timescale 35

2.4 Formal Assesment of Comparing Measured data and Modelled Data on a Daily timescale 41

2.4.1 Differences 41

2.4.2 Deming Regression Results 44

2.4.3 Peaks over Threshold Results 48

2.5 Conclusion 58

3 Spatial Analysis of the Monitoring Site, Diffusion Tube, DEFRA and ADMS-Urban Modelled Data **60**
3.1 Introduction 60
3.2 Methodology 64
3.2.1 Parameter Estimation 64
3.2.2 Variograms and Correlation Functions 65
3.2.3 Exploring the presence of spatial correlation in data 67
3.2.4 Spatial Prediction 68
3.3 Results 69
3.3.1 Exploring the differences predicted by the modelled and observed data 83
3.4 Conclusion 84

4 Investigating the characteristics of the ADMS-Urban modelled pixels in space **86**
4.1 Introduction 86
4.2 Methodology 89
4.2.1 Representing functions by basis functions 89
4.2.2 Choosing the number K of basis functions 92
4.2.3 Functional Principal Component Analysis 94
4.2.4 Clustering 96
4.3 Results 99
4.4 Conclusion 110

5 Functional calibration of the ADMS-Urban model output **112**
5.1 Introduction 112
5.2 Methodology 115
5.2.1 Estimation for the Concurrent Model 116
5.2.2 Missing Data 118
5.2.3 Selecting the Number of Basis Functions 119
5.3 Results 120
5.3.1 Monitoring Data 120
5.3.2 Diffusion Tube Data 124
5.4 Conclusion 126

6 Conclusion and Discussion **128**

6.1 Introduction . . . 128

6.2 Monitoring site and model comparison . . . 129

6.3 Spatial Comparison . . . 130

6.4 Dimension reduction and common behaviours in ADMS-Urban model . 131

6.5 Further Work . . . 132

Appendices **134**

A Daily Maximum and Daily Maximum Difference between the days plots **135**

B Map of the annual modelled prediction standard errors **138**

C Map of the monthly modelled prediction standard errors **139**

List of Tables

1.1 National air quality objectives and European Directive limit and target values for the protection of human health (DEFRA, 2007) 4
1.2 Monitoring sites and the pollutants they measure 11
1.3 Monitoring sites and their site classification 12
1.4 Sources of data and their temporal and spatial resolutions 14

2.1 Boundary layer variables computed by ADMS-Urban 19
2.2 Monitoring sites and the % of missing data 29
2.3 Mean difference, Variance difference and 95% Confidence Intervals (CI) for the mean difference 43
2.4 Summary of Linear Regression Model 44
2.5 Summary of Deming Regression Model 45
2.6 Summary of Linear Regression and Deming Regression Model 47
2.7 Number of exceedances over the 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles for both the modelled and monitoring data 49
2.8 Number of Exceedances over the 90^{th} percentile for both the modelled and monitoring data for months Jan to June over the year 2012 53
2.9 Number of Exceedances over the 90^{th} percentile for both the modelled and monitoring data for months July to Dec over the year 2012 54
2.10 Thresholds chosen for each of the six monitoring sites for both the modelled and monitoring data 57
2.11 $\hat{\lambda}$ values based on the thresholds chosen for each of the six monitoring sites for both the daily modelled and monitoring data 57

3.1 Results from linear model 71

4.1 Summary of the optimum average silhouette width for the zoomed in city centre including roads region of 7454 pixels 100

List of Figures

1.1 Plot of Locations of Monitoring Stations and Diffusion Tubes in Aberdeen City on a 1 km by 1 km grid 12

2.1 Plume rise model (CERC, 2013) 20

2.2 Plot of Locations of Monitoring Stations in Aberdeen on a 75 m x 75 m grid with gridded roads 21

2.3 Plot of NO_2 concentrations over Aberdeen at given time points (13^{th} December 2012 at 9 am (Winter) and 28^{th} June 2012 at 9 am (Summer)) 22

2.4 Monthly Mean Plots of NO_2 concentration at all six monitoring sites with 95% confidence bands represented by the bars (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 30

2.5 Monthly Maximum Plots of NO_2 concentration at all six monitoring sites (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 32

2.6 Monthly Mean Plot of NO_2 concentrations for diffusion tubes (Diffusion tube data is represented by the black line and the ADMS-Urban modelled data is represented by the red line). Vertical lines separate the data for each diffusion tube over the year. 34

2.7 Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Wellington Road (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 35

2.8 Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Union Street Roadside (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 36

2.9 Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Anderson Drive (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 36
2.10 Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Market Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 37
2.11 Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Errol Place (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 37
2.12 Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site King Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 38
2.13 Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Wellington Road (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 39
2.14 Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Union Street Roadside (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 40
2.15 Daily Mean Differences of the modelled data (μgm^{-3}) minus the monitoring data (μgm^{-3}) where the red line represents the horizontal line, $x = 0$ 41
2.16 Bland Altman plots (Daily Mean Differences against Daily Mean Averages) where the red line represents the mean difference, the blue lines represent the lower and upper 95% confidence intervals and the green line represents the horizontal line, $x = 0$ 42
2.17 Scatterplots of Modelled data (μgm^{-3}) against Monitoring data (μgm^{-3}) with linear and deming regression lines, 95% confidence intervals for the linear regression line and the line $y = x$ 46

2.18 Scatterplot of Modelled data (μgm^{-3}) against Diffusion Tube data (μgm^{-3}) with linear and deming regression lines, 95% confidence intervals for the linear regression and the line $y = x$ 47
2.19 Empirical CDF at all six monitoring sites (Modelled data) with horizontal lines representing the 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles and the green vertical line represents the value at which the 90^{th} percentile cuts the CDF 50
2.20 Empirical CDF at all six monitoring sites (Monitoring data) with horizontal lines representing the 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles and the green vertical line represents the value at which the 90^{th} percentile cuts the CDF 51
2.21 The time points at which the daily mean NO_2 modelled data and the daily mean NO_2 monitoring site data exceed the 90^{th} percentile 52
2.22 Mean Residual Life Plots for all six sites (ADMS-Urban modelled data). 55
2.23 Mean Residual Life Plots for all six sites (Monitoring site data). 56

3.1 Plot of Total Annual Mean concentrations for the pollutant NO_2 over the year 2012 in Aberdeen on 1 km by 1 km grids 62
3.2 Graphic representation of a classic variogram, with structural parameters specified (Diggle and Ribeiro Jr., 2007) 66
3.3 Plots to give an initial impression of what the data looks like 70
3.4 Q-Q Plot of $\log(NO_2)$ concentrations 71
3.5 Residual plots 72
3.6 Monte Carlo envelopes for the variogram of residuals after fitting the simple linear model 73
3.7 Map of the annual kriged observed $\log(NO_2)$ data and map of the observed prediction standard errors for 2012 73
3.8 Map of the annual kriged modelled $\log(NO_2)$ data 74
3.9 Map of the monthly kriged observed NO_2 data (January to June) . . . 76
3.10 Map of the monthly kriged observed NO_2 data (July to December) . . 77
3.11 Map of the observed prediction standard errors (January to June) . . . 78
3.12 Map of the observed prediction standard errors (July to December) . . 79
3.13 Map of the monthly kriged modelled NO_2 data (January to June) . . . 80
3.14 Map of the monthly kriged modelled NO_2 data (July to December) . . 81
3.15 Map of the modelled prediction standard errors (January and February) 82

3.16 Map of the model measurement differences 83
3.17 Monte Carlo envelopes for the variogram of the model measurement differences 84

4.1 Cumulative Variance Explained against Component for the zoomed in city centre region of 7454 pixels 100
4.2 Maps of Aberdeen highlighting the clusters using the partitioning around medoids clustering algorithm (Black circles represent the 6 monitoring site locations in Aberdeen and the yellow circles represent the diffusion tube locations) 101
4.3 Maps of Aberdeen highlighting the clusters in the summer (June, July and August) months and winter (January, February and March) months using the partitioning around medoids clustering algorithm (Black circles represent the 6 monitoring site locations in Aberdeen and the yellow circles represent the diffusion tube locations) 102
4.4 Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the main city centre of Aberdeen including the roads and monitoring daily NO_2 data for each site represented by the black line 104
4.5 Cluster mean curves for cluster 4 and cluster 5 represented by the purple and orange line respectively in the background pixels of Aberdeen and monitoring daily NO_2 data for each site represented by the black line . 105
4.6 Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the main city centre of Aberdeen in the summer months (June, July and August) including the roads and monitoring daily NO_2 data for each site represented by the black line 106
4.7 Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the background pixels of Aberdeen in the summer months (June, July and August) and monitoring daily NO_2 data for each site represented by the black line 107
4.8 Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the main city centre of Aberdeen in the winter months (January, February and March) including the roads and monitoring daily NO_2 data for each site represented by the black line . 108

4.9 Cluster mean curves for cluster 1, cluster 2 and cluster 3 represented by the red, blue and green line respectively in the background pixels of Aberdeen in the winter months (January, February and March) and monitoring daily NO_2 data for each site represented by the black line . 109

5.1 Plots of Errol Place with both missing values included and missing values imputed 119
5.2 Smoothed functions for modelled and monitoring data where 12, 24 and 36 basis functions have been used 121
5.3 Slope functions produced from running the concurrent functional linear model 122
5.4 Slope function with 95% confidence bands (12 basis functions) 123
5.5 Slope function with 95% confidence bands (24 basis functions) 123
5.6 Slope function with 95% confidence bands (36 basis functions) 124
5.7 Slope functions produced from running the concurrent functional linear model 125
5.8 Slope functions with 95% confidence bands produced from running the concurrent functional linear model 126

A.1 Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Anderson Drive (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 135
A.2 Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Market Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 136
A.3 Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Errol Place (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 136
A.4 Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site King Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line) 137

B.1 Map of the annual modelled standard errors (Including and Excluding the roads) 138

C.1 Map of the modelled prediction standard errors (March to June) 139

C.2 Map of the modelled prediction standard errors (July to December) . . 140

Acknowledgements

I would like to take this opportunity to thank my supervisor Marian Scott for her continued help and guidance throughout this thesis and for keeping me positive. I would also like to take this time to thank Alan Hills and the Airmod team at SEPA for their support and for funding my research. I would like to thank all my friends especially Francesca Pannullo for listening to my problems and making them all seem far easier.

Finally I would like to say a huge thank you to my Dad, Nana and my boyfriend Stephen for putting up with my stressful self and who will be more delighted than myself to hear that the masters is complete.

Declaration

This thesis has been completed by myself and no section of it has been submitted previously as part of any application for a degree. All work contained within this thesis was carried out by myself, except where otherwise stated.

Chapter 1

Introduction

1.1 Air Pollution and the Effects on Human Health

Air quality in the United Kingdom (UK) over the last ten years has improved. Since the industrial revolution, the air we breathe is overall purer today than any time prior to now (Department for Environment, Food and Rural Affairs (DEFRA), 2007). The Clean Air Act was launched in 1956 due to the Great Smog in London killing as many as four thousand people in 1952 (Environmental Audit Committee, 2011). The Clean Air Act was stated as "An Act to make provision for abating the pollution of the air" (Clean Air Act, 1956). Air pollution is produced, both through human activity (road traffic, industrial processes, combustion of oil and wood, construction machinery and shipping) and natural processes (volcanic eruptions and forest fires). The main human activity that contributes to air pollution is road traffic, since the emissions from road traffic happen at road level, hence humans are receptors (Stockholm, 2014). Air pollution produced especially from road traffic is thought to have a harmful effect on our health and wellbeing. DEFRA and The Rt Hon Elizabeth Truss MP (DEFRA, March 2014) states that there is growing confirmation of an association between road traffic and illnesses such as heart attacks and strokes. These sources also express that air pollution is calculated to have an effect equal to 29000 deaths every year and is likely on average to decrease the lifetime of every individual in the UK by six months at a value of about £16 billion per year (DEFRA, March 2014). The main air pollutants are Nitrogen Dioxide (NO_2), Nitric Oxide (NO), Ozone (O_3), Particulate Matter 10 (PM_{10}), Particulate Matter 2.5 ($PM_{2.5}$), Carbon Monoxide (CO) and Sulphur Dioxide (SO_2).

The Committee on the Medical Effects of Air Pollutants (COMEAP) describes the main air pollutants and its harmful effects to humans. COMEAP states that Particulate matter (PM) is the outcome of organic and inorganic substances. These particles can be primary or secondary, with primary indicating that the particles go directly into the atmosphere and secondary indicating that the particles are created due to a chemical reaction. It is thought that PM can worsen the illnesses of individuals suffering from heart disease, lung disease, bronchitis and individuals with asthma. SO_2 occurs as a gas but can be formed into an acidic solution when it is dissolved in water. This gas is produced when fuels containing sulphur are burned and can cause problems for people with asthma resulting in them having difficulty breathing. Moreover, during pollution periods there is potentially an increased risk of asthma attacks due to high levels of SO_2. NO_2 is a gas and is the outcome of the oxidation of NO by oxygen (O_2) or O_3 in the air. This gas can also be produced directly from vehicle exhausts. NO_2 can increase sensitivity to respiratory infections and to allergens due to its effects on the immune cells in the lungs. It can also act as an irritant at large concentrations which causes swelling of the airways (COMEAP, 2011).

Additionally, CO is a gas which is also harmful for humans. This gas has no colour, no smell and no taste and is produced when fossil fuels, wood and charcoal are burned without a sufficient supply of O_2. High concentrations of CO especially indoors can be fatal as it stops the normal transport of O_2 by the blood and also stops its transport to the body's tissues. Another air pollutant, O_3 is created through chemical reactions which are powered in the troposphere by sunlight. It is a secondary pollutant gas and is created by the effect of ultraviolet (UV) light on O_2 molecules in the stratosphere. At high levels of O_3, the eyes, nose and throat can be found to get irritated. This causes coughing and chest pain when breathing. When high pollution periods occur, high levels of O_3 may have an impact on individuals with asthma and like SO_2 may set off asthma attacks (COMEAP, 2011). The analysis in this thesis will focus on NO_2.

1.2 Current Air Pollution Guidelines

The Air Quality in Scotland (AQS) website states that air pollution standards are average levels over a specified time period that are thought to be suitable in terms of what is recognised about the effects that each of the pollutants have on health and the environment (AQS, 2014a). DEFRA (2011b) states that the European (EU) legislation is highly motivated in taking action to control and enhance air quality. The legally binding limits for concentrations in air outside of vital air pollutants that affect the health of the public are set by the 2008 ambient air quality directive (2008/50/EC). These vital air pollutants include PM_{10}, $PM_{2.5}$ and NO_2. Not only do these pollutants have direct impacts, they also can merge in the atmosphere to create O_3. O_3 is a dangerous air pollutant and a powerful greenhouse gas which can travel long distances due to weather systems (DEFRA, 2011b). Table 1.1 is taken from (DEFRA, 2007) and the vital air pollutants i.e NO_2, PM_{10} and $PM_{2.5}$ are observed in Table 1.1, for more information on the other pollutants see (DEFRA, 2007). From Table 1.1 the air quality objectives for PM_{10}, $PM_{2.5}$ and NO_2 for the UK and the Scotland specific objectives can be seen. The table shows that the UK 24 hour mean objective expresses that the pollutant PM_{10} should not exceed $50\mu gm^{-3}$ more than 35 times a year. Also the UK annual mean objective expresses that PM_{10} should not exceed $40\mu gm^{-3}$ and both of these targets have been in place in the UK since the 31^{st} December 2004. The targets for Scotland are different and it could be suggested that they are more firm. From the table it can be noted that the Scottish 24 hour mean expresses that PM_{10} should not exceed $50\mu gm^{-3}$ more than 7 times a year and the annual mean expresses that PM_{10} should not exceed $18\mu gm^{-3}$. Both of these targets have been in place in Scotland since 31^{st} December 2010. The table can be read in the same way for both $PM_{2.5}$ and NO_2.

Table 1.1: National air quality objectives and European Directive limit and target values for the protection of human health (DEFRA, 2007)

Pollutant	Applies	Objective	Concentration measured as	Date to be achieved by & maintained thereafter
PM_{10}	UK	$50 \mu g m^{-3}$ not to be exceeded more than 35 times a year	24 hour mean	31^{st} Dec 2004
PM_{10}	UK	$40 \mu g m^{-3}$	annual mean	31^{st} Dec 2004
PM_{10}	Scotland	$50 \mu g m^{-3}$ not to be exceeded more than 7 times a year	24 hour mean	31^{st} Dec 2010
PM_{10}	Scotland	$18 \mu g m^{-3}$	annual mean	31^{st} Dec 2010
$PM_{2.5}$ Exposure Reduction	UK (except Scotland)	$25 \mu g m^{-3}$	annual mean	2020
$PM_{2.5}$	Scotland	$12 \mu g m^{-3}$	annual mean	2020
NO_2	UK	$200 \mu g m^{-3}$ not to be exceeded more than 18 times a year	1 hour mean	31^{st} Dec 2005
NO_2	UK	$40 \mu g m^{-3}$	annual mean	31^{st} Dec 2005

1.3 The City of Aberdeen

This study focuses on the air quality in the city of Aberdeen. This section will go on to give a brief description of Aberdeen and the layout of the city. Aberdeen has 220,000 inhabitants and is located on the east coast of Scotland by the North Sea (Aberdeen City Council, 2013).

The city is well known for its services to the oil industry. Atmospheric pollution in the city is mainly caused by road traffic and it is explained by Aberdeen City Council (2013) that there are restricted ways of entering into the city or travelling around it. This is caused by the fact that the River Dee which lies to the south of Aberdeen and the River Don which lies to the north of Aberdeen restrict the road transportation system. By 2018, it is expected that the establishment of a Western Peripheral Route around the city will be finished. The main ways to converge or travel via the city centre are the A90 and A96 trunk roads, A93 North Deeside Road, A956 Ellon Road and A956 Wellington Road. Moreover, most of the traveller traffic that arrives into Aberdeen comes from Aberdeenshire (Aberdeen City Council, 2013).

Aberdeen Harbour is situated in the city centre and Aberdeen Airport (Dyce) is situated about 7 km to the northwest of the city of Aberdeen. The Harbour is an increasing environment serving as the UK's foremost base for supply ships and large boats to offshore establishments. Meanwhile, there are also regular ferries to The Shetland and Orkney Islands (Aberdeen City Council, 2013).

1.4 Air Quality Modelling

Air quality modelling is a crucial tool for expanding and assessing air quality policy (DEFRA, 2011a). In this section the ways in which air pollution are modelled is discussed with the inputs of the model and how the various modelling systems produce gridded maps being of particular interest. Our main focus is the air modelling for DEFRA and the Atmospheric Dispersion Modelling System (ADMS) models. The ADMS models are currently being used in a study in Aberdeen and these are favoured software of the Environmental Agency (EA) and other organisations (Mabbett, 2014). In particular, for this thesis, ADMS-Urban and the DEFRA Pollution Climate Mapping (PCM) models are of interest.

1.4.1 Air Modelling for DEFRA

There are six main models and many others that are presently used by DEFRA and the Devolved Administrations. These are Pollution Climate Mapping (PCM), Community Multi-scale Air Quality Modelling System (CMAQ), Fine Resolution Atmospheric Multi-pollutant Exchange (FRAME), European Monitoring and Evaluation Program Unified Model for the UK (EMEP4UK), Ozone Source Receptor Model (OSRM) and the UK integrated assessment model (UKIAM).

The PCM is a group of models that are aimed in such a way to satisfy part of the UK's European Union (EU) Directive (2008/50/EC) requirements to investigate and give results on the concentrations of certain pollutants in the atmosphere. Ricardo-AEA run these models on behalf of DEFRA and there is one model used for each pollutant. The pollutants considered are mono-nitrogen oxides NO_x, NO_2, PM_{10}, $PM_{2.5}$, SO_2, CO, benzene, O_3, Arsenic (As), Cadmium (Cd), Nickel (Ni), Lead (Pb) and Benzo(a)pyrene (B[a]p). All of the models are broken into two parts: the first of these parts is a base year model and then secondly a projections model. Outputs on a 1×1 km grid of background conditions are produced. Additionally about 9000 indicative road side values are produced. This model is used to construct background maps for the UK. These are 1×1 km grids of pollutant concentrations (DEFRA, 2013). These data can be downloaded from (DEFRA, February 2014). For more information on Ricardo-AEA see (Ricardo-AEA, 2013).

The CMAQ models are again operated by Ricardo-AEA on behalf of DEFRA. These models are used to compute daily air quality forecasts (DEFRA, 2013). These results can be seen on the DEFRA website (DEFRA, 2014a). The CMAQ is an open source model which was established by the US Environmental Protection Agency (USEPA). The CMAQ model outputs to a 50 $\times$ 50 km resolution over Europe. Within this output, there are 10 $\times$ 10 km squares for the UK. The FRAME model is a Lagrangian statistical trajectory model. It is operated by the Centre for Ecology and Hydrology (CEH) on behalf of DEFRA. For more information on CEH see (CEH, 2014). The aim of these models is to compute yearly averages of certain pollutants for both wet and dry deposition at a 5 $\times$ 5 km resolution. Furthermore, the FRAME model can also be used to compute high speed calculations that answer policy concerns and help debates through input to the UK Integrated Assessment Model (DEFRA, 2013). For more information on FRAME models see (DEFRA, 2009). The EMEP4UK produces evaluations of vital load exceedances. This model is again managed by CEH and an outer grid of 50 $\times$ 50 km^2 is utilized. Within this output, there are 5 $\times$ 5 km^2 squares for the UK. Through the development of an Eulerian multi-scale, multi pollutant model DEFRA will be able to establish the impact of distinct policy scenarios on various pollutants by only needing to run the model once (DEFRA, 2013).

Additionally, the OSRM model is used to give guidance on the impacts of arranged or suggested policy on O_3 concentrations to adjustments in precursor emissions. This model is operated by Ricardo-AEA and is a source-receptor Lagrangian trajectory model. This model has a single vertical layer that sits over the UK. Concentrations of O_3, NO and NO_2 are presented hourly in the mid-boundary layer at specified receptors. Results can also be presented on a 10 $\times$ 10 km UK grid. The UKIAM model is used to examine cost effective strategies for decreasing UK emissions. Subsequently, by examining cost effective strategies, developments in environmental protection in the UK are maximised while satisfying future UK emission ceilings enforced to decrease transboundary air pollution in Europe. This model gathers together information on various pollutants to compute the impact of abatement measures at the same time on a mixture of pollutants and the contrast of scenarios in the future (DEFRA, 2013).

1.4.2 Atmospheric Dispersion Modelling System

There are five types of the ADMS pollution models. These are the ADMS 5, ADMS-Urban, ADMS-Roads (Extra), ADMS-Airport and ADMS-Screen. The Cambridge Environmental Research Consultants (CERC) state that these models have been progressed in order to take advantage of the latest understanding of the way in which lower levels of the atmosphere behave in computer modelling systems for emissions into the atmosphere (CERC, 2014a). Each of the five models mentioned above focus on a different issue that air quality faces.

ADMS 5 is said to be the "**new generation Gaussian plume air dispersion model**" and models the effect of occurring and suggested industrial installations. This indicates that the atmospheric boundary layer properties are not just dependent on the single parameter Pasquill-Gifford class but instead are distinguished by two parameters namely the boundary layer depth and the Monin-Obukhov length (CERC, 2014b). ADMS-Urban is a wide-ranging modelling tool for addressing issues in air pollution. This model focuses on issues in huge urban areas, cities and towns (CERC, 2014c). ADMS-Roads is a wide-ranging modelling tool for looking into difficulties in air pollution caused by small networks of roads that may be associated with industrial sites. Additionally, ADMS-Roads Extra is an extension of the ADMS-Roads pollution model and generally at the same time lets additional sources be studied (CERC, 2014d). Another model is the ADMS-Airport pollution model which is an extensive tool used for controlling air quality at airports (CERC, 2014e). The last of the ADMS pollution models is the ADMS-Screen pollution model which is a screening model for computations involving air quality (for more information see CERC, 2014f). As mentioned previously, the model that will be focused on in this study is ADMS-Urban.

The model input data and model output data is now described for the ADMS-Urban pollution model. Information on the model input and model output data for the other ADMS models can be found on the CERC website (CERC, 2014a). These models overlap in some cases in terms of the model input and model output data. The ADMS-Urban pollution model requires a number of data to be input to the model. These include emissions sources, emissions profiles, meteorological data, traffic flow, background ambient concentrations and aggregated emissions. The ADMS-Urban pollution model can be used to explore emissions from as many as 7500 sources at the same time. These include road traffic where over 145000 road links can be modelled and 3000 road sources allowing all road sources to have up to 50 vertices, industrial sources where as many as 1500 point sources, line sources, area sources or volume sources can be inputted and aggregated sources. If emissions from sources are not big enough to be determined clearly then as many as 3000 grid cells may be utilized to model these (CERC, 2014g).

To account for the diurnal variation in traffic flows there can be as many as 500 user defined emissions profiles incorporated in any run of the model. Other emission profiles that can be incorporated are seasonal variations with as many as 500 monthly profiles, variation of sources with the way in which the wind is moving and as many as 500 yearly hourly profiles. A variation of meteorological data may be utilized for input and the way in which this data is inputted is straightforward. The meteorological data that are needed are wind speed, the direction of the wind and temperature. These can be inputted along with cloud cover, heat flux or solar radiation. The required boundary layer parameters are computed by the meteorological pre-processor using the input of the user. When modelling road sources, hourly speed and traffic flow data can be input into the model and ADMS-Urban's built-in emission factors is used. Roads in urban areas are much more complicated to model than traffic emissions as a line source. When modelling roads in ADMS-Urban both the impact of street canyons and turbulence caused by traffic are incorporated into the model. When any local emissions are modelled, it is vital to incorporate the background ambient concentrations that are advected outwith the area being modelled. These concentrations may be hourly values or assumed constant values if hourly values are not obtainable. From the DEFRA website http://uk-air.defra.gov.uk/data/ these background data can be downloaded for the UK. These background data can be put straight into any ADMS-Urban model no matter what the scenario is. As briefly mentioned at the end of the previous paragraph it is also crucial in urban areas to incorporate the aggregated emissions from sources that may not be big enough to be clearly determined but add to pollution levels overall. A grid source with as much as 3000 grid cells may be utilized to model these (CERC, 2014g).

From the model output, pollution concentrations can be computed for averaging times.

These can vary from seconds to as much as years. Computations for rolling averages, the amount of exceedances of threshold concentrations and percentile statistics are also available from the model output when using ADMS-Urban. The results from the model are generally firstly confirmed by comparing them with locally monitored data. This may be achieved by looking at the results outputted at receptor points that correspond to locations of monitoring sites and then the modelled and monitored concentrations can be plotted in a time series plot and compared. This thesis will explore similar time series plots in Chapter 2. Colour contour plots are usually used to present the output of the model. The intelligent gridding allows for modelling of areas which are larger and high spatial resolution is achievable in areas of specific interest for example in and around the roads (CERC, 2014g). The technical detail of the ADMS-Urban model is described in more depth in Chapter 2.

1.5 Description of the data

In this study four main sources of data (two measured and two modelled) are explored. These are data from the Air Quality in Scotland website (AQS, 2014b), data from DEFRA which can be found on the DEFRA website (DEFRA, 2014b), diffusion tubes and ADMS-Urban modelled data provided by the Scottish Environment Protection Agency (SEPA).

The data from the AQS website gives hourly measured pollutant levels at the six monitoring sites in Aberdeen. All six of these monitoring sites are Automatic Urban and Rural Networks (AURN). The AURN is the leading network utilised for compliance against the Ambient Air Quality Directives and it is also the biggest automatic monitoring network in the UK. It incorporates automatic air quality monitoring stations. These monitoring stations measure the pollutants nitrogen oxide (NO_x), sulphur dioxide (SO_2), ozone (O_3), carbon monoxide (CO) and particulate matter 10 and 2.5 (PM_{10}, $PM_{2.5}$) and the pollutants are measured hourly, providing high resolution data. Data for these are available for the public to download with various methods of doing so and the AQS website is one of those (DEFRA, 2012). The concentrations for each pollutant are measured in μgm^{-3}. The pollutants measured at the Aberdeen monitoring sites include nitrogen dioxide (NO_2), PM_{10}, $PM_{2.5}$, O_3, CO and SO_2. Table 1.2 below highlights the pollutants that each monitoring station measures in Aberdeen.

This information can be found on the AQS website (AQS, 2015).

Table 1.2: Monitoring sites and the pollutants they measure

Monitoring Site	Pollutants Measured
Union Street Roadside	NO_2, PM_{10}, $PM_{2.5}$
Anderson Drive	NO_2, PM_{10}
Errol Place	NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, SO_2
King Street	NO_2, PM_{10}
Market Street	NO_2, PM_{10}
Wellington Road	NO_2, PM_{10}

The data contains information from as far back as 1999 for the city of Aberdeen at the Errol Place monitoring site. However, it should be noted that the data at Anderson Drive, Market Street and Union Street Roadside monitoring sites goes as far back as 2005 and that the data at Market Street is recorded until January 2009 and then the monitoring site is replaced with Market Street 2. The Wellington Road monitoring site has air quality data that goes back as far as 2008 whereas the King Street and Market Street 2 monitoring sites only have data from 2009 onwards.

The AQS website (AQS, 2014c) describes the description, source and objectives of twelve different monitoring site classifications including other. In Aberdeen the six monitoring sites are all based around the city centre. All six of these sites are either classified as Roadside or Urban Background. A summary of the monitoring sites and their classifications are shown in Table 1.3. Sites that are classified as roadside are sites which sample "between 1 m of the kerbside of a busy road and the back of the pavement." This is usually within 5 m to 15 m of the road. Sites that are classified as urban background are sites "distanced from sources and therefore broadly representative of city-wide background conditions" (for more information see AQS, 2014c).

Table 1.3: Monitoring sites and their site classification

Monitoring Site	Site Classification
Union Street Roadside	Roadside
Anderson Drive	Roadside
Errol Place	Urban Background
King Street	Roadside
Market Street	Roadside
Wellington Road	Roadside

The data from DEFRA are background maps which are established from the year 2011. Data are available from years 2011 to 2030 for various pollutants. These data are produced through running the Pollution Climate Mapping (PCM) model. These maps are provided 1 km $\times$ 1 km grids (DEFRA, 2014b). The pollutant concentrations are total annual mean concentrations which have been established from 1 km $\times$ 1 km grid squares. Again the pollutant is measured in μgm^{-3}. An example of the DEFRA 1 km $\times$ 1 km grid and the locations of the monitoring sites in Aberdeen are illustrated in Figure 1.1. From Figure 1.1 it can be observed that spatially all of the monitoring stations are close to one another indicating that the level of pollution at each of these sites may all be somewhat similar.

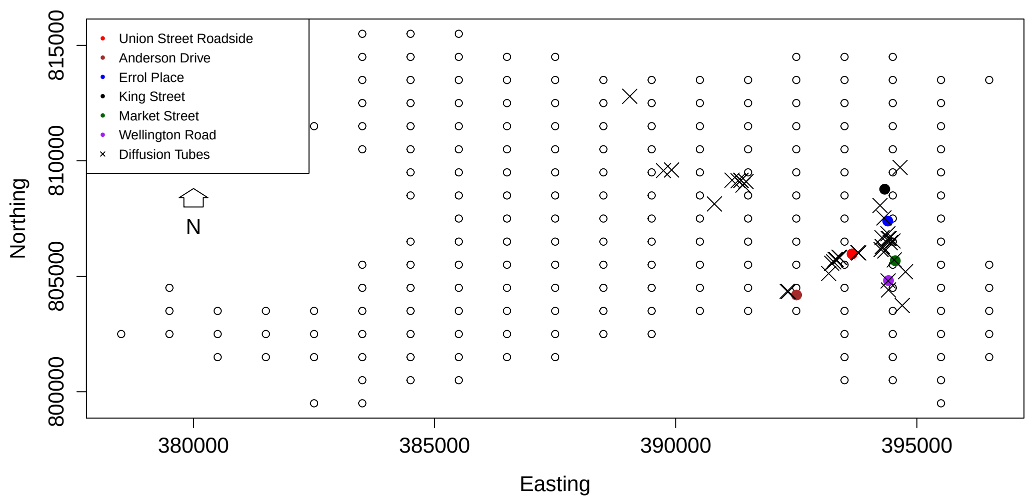


Figure 1.1: Plot of Locations of Monitoring Stations and Diffusion Tubes in Aberdeen City on a 1 km by 1 km grid

The diffusion tube data provided by SEPA are measured at an additional 46 locations which are also illustrated in Figure 1.1. These are passive samplers and are made up of miniature plastic tubes which include a chemical reagent. The pollutant of interest is soaked up by this chemical reagent. When considering Palmes-type NO_2 diffusion tubes, which is of interest for this thesis, the chemical reagent used is triethanolamine (TEA). A water or acetone based solution of this chemical reagent covers the stainless steel mesh grids at the end of the tube where it closes (DEFRA, 2008). These diffusion tubes are connected to lampposts and downpipes and strongly suggest longer-term average NO_2 concentrations. They also point out and indicate areas in Aberdeen where NO_2 concentrations are higher (Aberdeen City Council, 2013). These data are measured monthly and annual mean values have also been provided and again the pollutant concentrations are measured in μgm^{-3}. The monthly values are not exactly collected monthly i.e there is not an average pollutant concentration for January, February, March etc. Instead these values have been collected over 4/5 week durations but are not gathered at the beginning and end of the month. From Figure 1.1 it can be seen that the diffusion tube locations are spatially quite close to one another and around the monitoring site locations. There is however, diffusion tubes located near the northwest of the city of Aberdeen covering the airport.

The final main source of data was produced by running the ADMS-Urban model and these data were provided by SEPA. These data are background maps of the city centre of Aberdeen which indicate the roads in the city centre of Aberdeen. These maps are provided in 75 m $\times$ 75 m grids with gridded roads and there are 18319 gridded points in total. Throughout this thesis all the analysis will be focused on 2012. At each of these points concentration levels for pollutants, NO_2, mono-nitrogen oxides NO_x, PM_{10} and $PM_{2.5}$ have been measured every hour of the day over the year 2012. The pollutants were measured in μgm^{-3}. Moreover, the direction of the wind, 10 m wind speed and the height of the boundary layer divided by the monin-obukhov length were measured for every hour of the day over the year 2012.

These four main sources of data are of very different temporal and spatial resolutions with data measured hourly, daily, monthly and annually. Table 1.4 summarises each source of data and their temporal resolutions.

Table 1.4: Sources of data and their temporal and spatial resolutions

Data Source	Temporal Resolution	Spatial Resolution
AQS Data	Hourly, Daily, Monthly, Annually	Point
Diffusion Tube	Monthly, Annually	Point
DEFRA	Annually	1 km $\times$ 1 km grid
ADMS-Urban Model	Hourly, Daily, Monthly, Annually	75 m $\times$ 75 m grid with gridded roads

Spatially data has been collected at selected points and these data are given by the monitoring stations and the diffusion tubes in Aberdeen. Data have also been collected at 1 km by 1 km grids and these data are given by the background maps produced by DEFRA. Moreover, data has also been collected at 75 m $\times$ 75 m grids with gridded roads and these data are produced by running the ADMS-Urban model.

1.6 Aims

The following aims of this study are:

- Firstly, this thesis will investigate for each monitoring site point and overall diffusion tube data, how comparable the ADMS-Urban model and observed data are. This will give evidence of how well calibrated the observations and predictions are. However, it will also help indicate, given a particular threshold, whether the observations and predictions exceed that given threshold at the same points in time.

- The second aim is to then explore how comparable the ADMS-Urban model and observed data are over the full domain of Aberdeen.

- Finally, this thesis will look at functional data analysis to investigate the characteristics of the ADMS-Urban modelled pixels in space and also to see in a functional context how comparable the ADMS-Urban model and observed data are.

In the next chapter, Model Measurement Comparisons, it is of interest to investigate how comparable the ADMS-Urban model and observed data are. Firstly, monthly mean and maximum NO_2 concentration plots will be produced followed by daily mean and maximum NO_2 concentrations plots. Following this, more formal approaches such as errors in variables regression and peaks over threshold will be explored in detail to assess how well the modelled and observed data are calibrated.

In Chapter 3, the focus will move on to spatial analysis of the monitoring site, diffusion tube, DEFRA and ADMS-Urban Modelled data. The full region of Aberdeen will be investigated in this chapter and the aim will be to find out if over a spatial region the modelled and observed data are well calibrated. In order to do this, a statistical spatial model will be fitted and ordinary kriging will be carried out to produce a spatial surface of the predicted NO_2 concentrations. Chapter 4 will then move on to investigate the characteristics of the ADMS-Urban modelled pixels in space. In this chapter, functional data analysis will be explored and more specifically functional principal components analysis (PCA) will be carried out and the principal components scores will be used in order to carry out functional clustering.

Chapter 5 will then investigate functional calibration of the ADMS-Urban model output through using techniques such as fitting functional linear models to the modelled and observed data. In this chapter we will explore in a functional context how well the modelled and observed data are calibrated. Finally Chapter 6 will conclude and discuss what has been found throughout this thesis and discuss what could be done as further work.

Chapter 2

Model Measurement Comparisons

In this chapter the ADMS-Urban model will be explored and analysed to see how well this model compares with observed NO_2 concentrations in Aberdeen. The predictions from the model will be compared to the monitoring site data and the diffusion tube data which will be considered as measured data. This chapter will look at how well the modelled and monitoring data are calibrated and then go on to look at extreme values to see if they occur at the same points in time and for this piece of analysis the monitoring site data will be focused on.

Righi *et. al* (2009) in their paper discuss the comparison of carbon monoxide (CO) concentrations estimated by the ADMS-Urban air quality forecasting model with concentrations that have been measured by air quality monitoring network stations. The study was carried out in the urban area, Ravenna (North East Italy). Two datasets were taken into consideration in their analysis. The first dataset was given by the mass-consistent meteorological pre-processor CALMET, this utilises data from Northern Italy's surface and upper air stations. For more information see (Holtslag and Van Ulden, 1983; Scire *et. al*, 2000; Deserti *et. al*, 2011). The second dataset was calculated based on a meteorological station situated in the city centre of Ravenna. Two monitoring sites situated in the city centre of Ravenna were used and hourly data was used to evaluate the potential of the model. Indexes such as Pearson's product-moment correlation coefficient (COR), normalised mean square error (NMSE), factor two (FA2), fractional bias (FB), index of agreement (IA) and the factor of exceedance (FOEX) were used in order to determine how well the model performed. For more information on these indexes see Righi *et. al* (2009), also for more information on IA and FOEX, see Wilmott (1982) and Sokhi *et. al* (2005) respectively.

Righi *et. al* (2009) were interested in investigating the dissimilarity between the predictions made by the model and the data measured in terms of atmospheric stability and wind speed and direction. In this study Righi *et. al* (2009) compared a meteorological dataset given by a pre-processor with a dataset achieved from measures calculated at a meteorological site situated in Ravennas city centre. This research highlighted that CO concentration measurements predicted by using the meteorological dataset are similar to measurements calculated at monitoring sites. The ADMS-Urban model seems to perform adequately in predicting CO concentrations although in the case under study it is found that the ADMS-Urban model tends to under estimate CO concentrations than the measured data.

Righi *et. al* (2009) investigated to see what could make the model perform better and they concluded that by adding a correction to the predicted values, via the evaluation of the running mean helps improve model performances in this study. Then the concentration data which was simulated is adjusted by taking the hourly average concentration that has been simulated and adding to it the difference between the hourly average that has been evaluated and the appropriate running average. Finally, it was concluded that wind speed plays an important role in model performance when the wind speed is < 4 m s^{-1} and that wind direction seems to play a significant role in model performance only when wind speed is again < 4 m s^{-1} (Righi *et. al*, 2009).

Oberkampf and Barone (2006) discuss in their paper several characteristics that they consider should be included and excluded in a validation metric. They describe a validation metric as measures that can be calculated that can quantitatively contrast both computational and experimental results across a span of variables (both input or control) to enhance evaluation of computational precision. Oberkampf and Barone (2006) use confidence intervals to establish a new validation metric. They build one particular metric requiring interpolation of experimental data and one particular metric that needs regression of experimental data. In this analysis these metrics are implemented to three different examples and Oberkampf and Barone (2006) examine how these metrics can be simply understood for evaluating precision of computational models. They also explore how these metrics can be simply understood for the affect of experimental measurement uncertainty on the precision evaluation (Oberkampf and Barone, 2006).

Bennett *et. al* (2012) analyse methods that are accessible across several fields for distinguishing the way in which environmental models behave. The main focus in their paper is numerical, graphical and qualitative procedures. They discuss many things within their paper including "general classes of direct value comparison, coupling real and modelled values, preserving data patterns, indirect metrics based on parameter values, and data transformations." Bennett *et. al* (2012) put forward a process for assessing model performance; the first step involves reassessing the scope, scale and aim of the model. The second step of the process requires describing the features of the data for calibration and testing. The third step then entails observing and using other ways of examination to identify the behaviours of under-modelled or non-modelled and to obtain a general summary of performance where everything has been taken into account. The fourth step involves choosing criteria for primary performance. The final step of the process involves examination of techniques that are more advanced to take care of issues such as standard differences between modelled and observed values (Bennett *et. al*, 2012).

2.1 Model Description

This section will describe the ADMS-Urban model in more technical detail and also give insight to the set up of the model used in this analysis.

2.1.1 ADMS-Urban Model

This section is based on information given by CERC (2013). As previously mentioned in Section 1.4.2, ADMS-Urban is a wide-ranging modelling tool for addressing issues in air pollution and focuses on issues in huge urban areas, cities and towns (CERC, 2014c). For the ADMS-Urban model, wind speed and wind direction are two of the meteorological variables that must be included. As well as these two variables, one of the following must also be included: the reciprocal of Monin-Obukhov length, the surface sensible heat flux or the cloud cover, time of day and time of year. It is preferable if a sensible approximation is known to include the variables temperature and boundary layer height if only cloud cover, time of day and time of year have been defined.

The boundary layer height h and the Monin-Obukhov length L_{MO} specify the boundary layer in ADMS-Urban. The Monin-Obukhov length is given by:

$$L_{MO} = \frac{-u_*^3}{\left(\frac{kgF_{\theta 0}}{\rho c_p T_0}\right)}. \tag{2.1}$$

From Equation 2.1:

- u_* denotes the friction velocity at the Earth's surface,
- k denotes the von Karman constant (0.4),
- g denotes the acceleration due to gravity,
- $F_{\theta 0}$ denotes the surface sensible heat flux,
- ρ denotes the density heat capacity of air,
- c_p denotes the specific heat capacity of air and
- T_0 denotes the near-surface temperature.

The Monin-Obukhov length is either negative or positive depending on the stability of the conditions. It is negative in unstable conditions and positive in stable conditions. Table 2.1 highlights the boundary layer variables that the ADMS-Urban computes. These are computed at various heights and vertical profiles are indicated as two functions namely z/L_{MO} and z/h where z denotes the height above the ground.

Table 2.1: Boundary layer variables computed by ADMS-Urban

Variable	Description
$U(z)$, $\frac{dU}{dz}$, $\frac{d^2U}{dz^2}$	Mean wind speed (m s^{-1}) and its first (s^{-1}) and second derivatives with height (m^{-1} s^{-1})
$\sigma_u(z)$, $\sigma_v(z)$, $\sigma_w(z)$	Root mean square turbulent velocities (m s^{-1})
$\Lambda_v(z)$, $\Lambda_w(z)$	Turbulent length scales (m)
$\varepsilon(z)$	Energy dissipation rate (m^2 s^{-3})
$T_L(z)$	Lagrangian time scale (s)
$N(z)$	Buoyancy frequency (s^{-1})
$T(z)$	Temperature (K)
$\rho(z)$	Density (kg m^{-3})
$P(z)$	Pressure (mbar)

These boundary layer variables are utilized to compute σ_y and σ_z which denote the plume spread parameters. Therefore, the plume spread parameters differ with plume and source height. It should also be noted that the turbulent velocities σ_u, σ_v and σ_w tend to have a small value of turbulence. This can be anywhere between 0 and 0.2 m s^{-1} and is determined by the minimum value the user inputs for L_{MO}. This takes into consideration that in urban areas where conditions are usually unstable, it is certain that there will be some turbulence.

The rise in trajectory and intensified dilution of a continuous discharge of gaseous substance that has a high temperature or momentum is estimated by the plume rise module. When producing plume rise, it takes into consideration the effects of plume buoyancy and momentum and incorporates penetration of inversions.

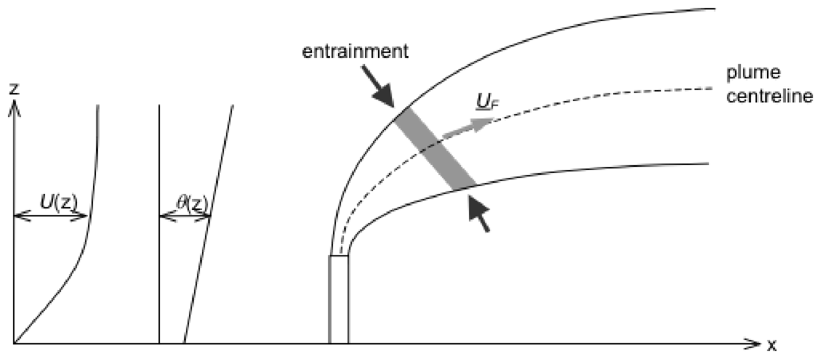


Figure 2.1: Plume rise model (CERC, 2013)

The source conditions exit diameter, emission velocity or volume flow rate and temperature or density determine the starting values of the plume rise module. The solutions to the equations are then worked out numerically using a Runge-Kutta numerical scheme with a changeable internal time step (CERC, 2013).

The model run given by SEPA for this thesis was set up with 181 road sources and no industrial or grid sources. The chemistry module has been turned on as NO_x chemistry ensures that precise predictions of NO_2 concentrations are produced. Buildings was turned off as this only works for point sources and not road sources, complex terrain was also turned off. Details of the roads have also been incorporated, these include width of the road, canyon height and elevation of the road, where the elevation of the road was always zero. Time varying emission factors were also taken into consideration within the model run, for example, weekdays may be different from Saturday and

Sunday. For the background concentrations, observed data from Errol Place was used which suggests that the model may appear to perform better at Errol Place. The grid spacing from the output of the model is regular. However, the roads are dealt with differently, and the grids are not equidistance.

An example of the ADMS-Urban 75 m × 75 m grid with gridded roads and the locations of the monitoring sites in Aberdeen are illustrated in Figure 2.2. From Figure 2.2, the classification of the monitoring sites are highlighted as it can be seen that all of the monitoring sites except Errol Place are Roadside. Once more it can be noted that spatially all of the monitoring sites are close to one another.

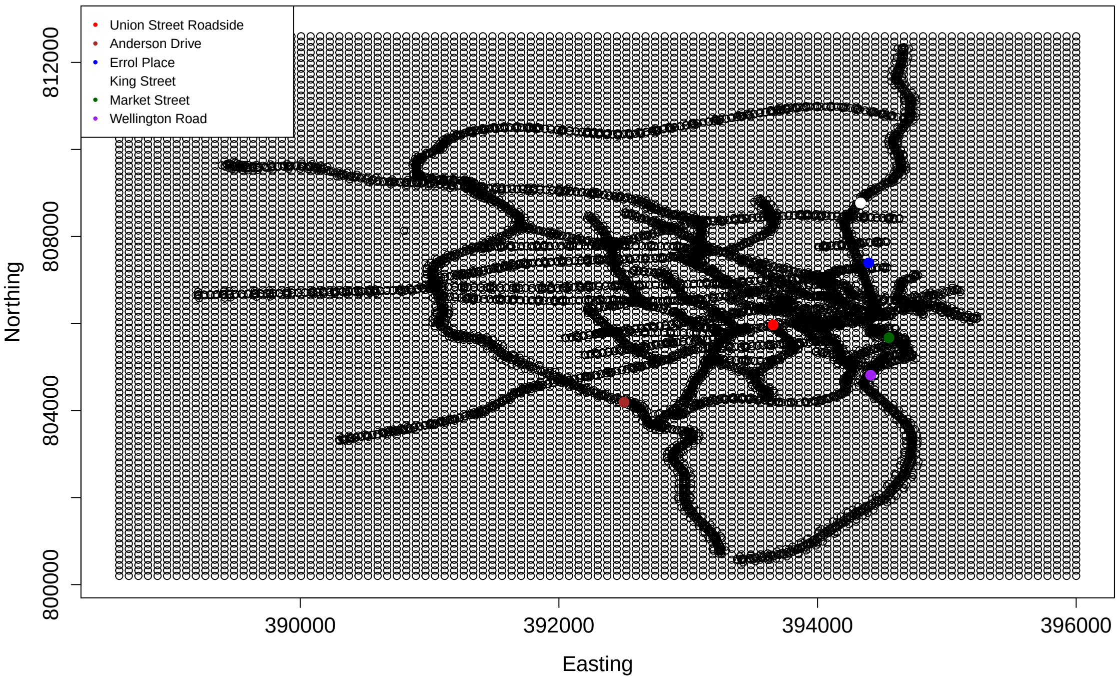
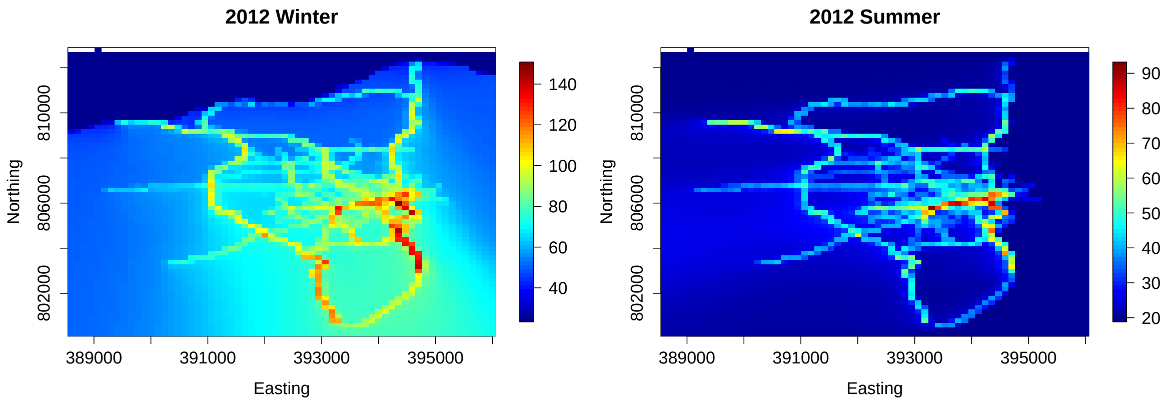


Figure 2.2: Plot of Locations of Monitoring Stations in Aberdeen on a 75 m x 75 m grid with gridded roads

Given particular time points throughout the year, for example 12^{th} June 2012 at 5pm the modelled air pollution concentrations over Aberdeen can be examined. Here it is of particular interest to consider a summer time point and a winter time point to see if there are any differences or similarities between both. The pollutant NO_2 will be looked at.



(a) Winter time point

(b) Summer time point

Figure 2.3: Plot of NO_2 concentrations over Aberdeen at given time points (13^{th} December 2012 at 9 am (Winter) and 28^{th} June 2012 at 9 am (Summer))

Looking at Figure 2.3 it can be observed that the modelled NO_2 concentrations for a particular winter day at 9 am are significantly higher than a particular summer day at 9 am. This is highlighted through looking at both of the scales in Figures 2.3a and 2.3b and noticing that the winter time point scale goes as high as 140 μgm^{-3} where as the summer time point scale goes as high as 90 μgm^{-3}. Furthermore, exploring Figure 2.3b it can be observed that most of the gridded points appear to be blue in colour indicating that the air pollution is low there and the points along the road appear to have higher levels of NO_2. Whereas investigating Figure 2.3a it is noticed that most of the gridded points appear to lie in the mid-range of the scale, around 60 to 100 μgm^{-3} with points along the roads even higher than this.

2.2 Methodology

This section will describe the formal assessments carried out in order to determine how well the ADMS-Urban model is performing in terms of its comparison to the monitoring data. Before describing these formal methods, some of the notation used throughout this section will be explained. Let X_i and Y_i denote the true values of the random variables where $i = 1, \ldots, n$ where n is the total number of observations and let x_i and y_i denote the observed values. The fitted values and the mean values of X_i are denoted by $\hat{X}_i$ and $\bar{X}_i$ respectively and the same would apply for Y_i. Also if a function t is considered then $\{t\}_+$ represents the positive part of t. Two methods will be used, the first method used is errors in variables (EIV) regression and the second method used is peak over threshold (POT) extreme value analysis.

2.2.1 EIV Regression

The aim of EIV regression is to consider the existence of error in both variables used to approximate the regression line. EIV regression is carried out here as there is clearly uncertainty in both sets of measurements (modelled and monitoring data) and carrying out standard regression doesn't account for the uncertainty in the explanatory variable. Gillard (2010) discusses in his paper that EIV regression has various applications these include, economic literature, astrostatistics, fisheries statistics and medical statistics. These are only a few of the many applications. The difficulty of errors in variables is fundamentally important in many applications and several theoretical progressions made in errors in variable regression have been constructed in their own right.

Let's suppose there is two variables say X and Y and the relationship between these two variables is linear and takes the form,

$$Y_i = \alpha + \beta X_i, \quad \text{where} \quad i = 1, \ldots, n. \tag{2.2}$$

From Equation 2.2, α denotes the intercept and β denotes the regression parameter. In this thesis, here Y_i would represent the ADMS-Urban modelled data and X_i would represent the monitoring site real data. Rather than observing the variables X_i and Y_i, the following

$$x_i = X_i + \delta_i, \tag{2.3}$$

$$y_i = Y_i + \varepsilon_i = \alpha + \beta x_i + \varepsilon_i, \tag{2.4}$$

are observed. From Equations 2.3 and 2.4 the random error elements are denoted by δ_i and ε_i. The expectation and variance of the random error elements are

$$\mathbb{E}(\delta_i) = \mathbb{E}(\varepsilon_i) = 0, \quad \forall\, i,$$

$$\mathrm{Var}(\delta_i) = \sigma_\delta^2, \quad \mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2, \quad \forall\, i.$$

It is also supposed that the both of the errors δ_i and ε_i are mutually uncorrelated. Therefore

$$\mathrm{Cov}(\delta_i, \delta_j) = \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j,$$

$$\mathrm{Cov}(\delta_i, \varepsilon_j) = 0, \quad \forall\, i, j.$$

Equation 2.4 can be rewritten as follows

$$y_i = \alpha + \beta x_i + (\varepsilon_i - \beta\delta_i), \quad \text{where} \quad i = 1, \ldots, n. \tag{2.5}$$

This emphasises the differences between standard regression modelling and errors in variable regression modelling. It can be clearly seen from Equation 2.5 that the error term depends on β and also that the term (ε - $\beta\delta$) depends on x. The covariance between (ε - $\beta\delta$) and x is given by

$$\text{Cov}(x, \varepsilon - \beta\delta) = \mathbb{E}[x(\varepsilon - \beta\delta)] = \mathbb{E}[(\xi + \delta)(\varepsilon - \beta\delta)] = -\beta\sigma_\delta^2,$$

where Cov(x, $\varepsilon - \beta\delta$) $= 0$ only if $\beta = 0$ or $\sigma_\delta^2 = 0$.

Two general classifications namely the functional model and the structural model can be used to divide the errors in variable modelling. The way in which the ξ_i are handled is the underlying difference between both of these models. The fundamental model supposes that the ξ_i's are not specified but fixed constants μ_i and the structural model supposes that the ξ_i's are a random sample from a random variable. This random variable has mean and variance μ and σ^2 respectively (Gillard, 2010).

Deming Regression

In this analysis Deming regression (Linnet, 1990) was carried out. Linnet states that here the error variances are assumed to be constant and their ratio known. The ratio is given by

$$\lambda = \frac{\sigma_\delta^2}{\sigma_\varepsilon^2}.$$

Applying a least squares method, the sum of squares to be minimised is given by

$$\sum_{i=1}^{n}[(x_i - \hat{X}_i)^2 + \lambda(y_i - \hat{Y}_i)^2],$$

where $\hat{Y}_i = \alpha + \beta(\hat{X}_i - \bar{X})$. For a solution to be achieved, the following sums are calculated for all n pairs of observations (x_i, y_i):

$$u = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$
$$q = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$
$$p = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

The slope and intercept estimates can be shown to be,

$$\hat{\beta} = \frac{(\lambda q - u) + \sqrt{[(u - \lambda q)^2 + 4\lambda p^2]}}{2\lambda p},$$

$$\hat{\alpha} = \bar{y} \text{ (Linnet, 1990).}$$

2.2.2 Bland Altman Plots

It is stated by Bland and Altman (1986) that for a new measurement approach to replace an old existing one in clinical measurement, comparison is frequently required to see whether they are adequately consistent. They suggest that the amount measurements can differ without creating issues is a matter of discernment, and preferably it should be described beforehand to assist in both the examination of contrasting the methods and selecting the sample size. They refer to Bland Altman plots as "a plot of the difference between the methods against their mean" and put forward that this may be an instructive approach. Any plausible association between the measurement error and the true value can be explored using these plots (Bland and Altman, 1986).

When there is not a clear relationship seen by observing the plot (i.e between the difference and the mean), the lack of agreement can be outlined by computing the bias. The bias is evaluated by the mean difference represented by $\bar{d}$ and the standard deviation (s.d.) of the differences represented by s. If the bias is unchanging in nature, it can be altered for by taking away the mean difference from the new measurement approach. If the differences follow a normal distribution which they most probable will, the majority (95%) of the differences are believed to lie between the limits $\bar{d} - 2s$ and $\bar{d} + 2s$. By producing a histogram of the differences, it can be examined whether the differences are normally distributed. Given the differences lie within $\bar{d} - 2s$ and $\bar{d} + 2s$ which will be referred to as the "limits if agreement", it can be concluded that both the new measurement approach and the old existing one can be used interchangeably. In this analysis, the two approaches are namely the modelled and monitored data. By producing these plots, the level of agreement between both sets of data can be observed (Bland and Altman, 1986).

2.2.3 Extreme Value Theory

The purpose of carrying out an extreme value analysis is to investigate whether the extreme events over a particular threshold occur at the same points in time in the modelled and measured data. The following idea described in this section is based on Coles and Davison (2008). Let $X_1, \ldots, X_n$ denote independent identically distributed random variables and let F denote the distribution that they follow and consider the straightforward case, $X_1, \ldots, X_n \sim F$. Precise inferences on tail of F are needed and there are three main problems. These are, that in the tail of the distribution there are few measurements, evaluations are frequently needed past the greatest measured data value and where the data have substantial density usual density estimation methods fit well but when evaluating tail probabilities can be extremely biased. Extreme value theory is widely known as having two fundamental application areas namely environmental and reliability modelling (Coles and Davison, 2008).

Point Process Approach and Peaks over Threshold

Peaks over thresholds is a particular instance of a point process representation. A number of points for example locations of stars in the sky are known as a point process which will be denoted by $\mathcal{P}$. For some acceptable set $\mathcal{A}$, the number of points occurring in $\mathcal{A}$ can easily be computed, and for some n which is random, the process can be written as

$$\mathcal{P} = \sum_{j=1}^{n} \delta_{X_j}$$

where the locations of the points are represented by the X_j and δ_x puts unit mass at x. Poisson process is the fundamental point process. For peaks over threshold modelling (POT), again it is supposed that there is a time series of observations and u which denotes the threshold that has to be determined. Determining the threshold u is a practical issue. The usual distributions used here are Pareto, Beta and Exponential which are all obtained from the Generalised Pareto distribution (GPD) for the exceedances (Coles and Davison, 2008).

Suppose $X^*_{n,i}$ = (X_i - b_n) / a_n, for i = 1, ..., n. Then the Poisson limit which is fully described by Coles and Davison (2008) states

$$\mathbb{P}\{X^*_{n,i} > u + x | X^*_{n,i} > u\} \approx \left\{1 + \xi \frac{x}{\sigma + \xi(u - \mu)}\right\}_+^{-1/\xi}.$$

Absorbing the scaling coefficients that are not known steers towards the survivor function of the GPD. This is given by

$$\mathbb{P}\{X^*_{n,i} > u + x | X^*_{n,i} > u\} = \left(1 + \xi \frac{x}{\tau}\right)_+^{-1/\xi}, \quad x > 0. \tag{2.6}$$

From Equation 2.6, $\tau = \sigma + \xi(u - \mu)$. The mean residual life plot is a way of identifying the "correct" threshold u. Given $\xi < 1$, the mean residual life occurs and fulfils

$$\mathbb{E}(X - u | X > u) = \frac{\sigma + \xi u}{1 - \xi}. \tag{2.7}$$

Equation 2.7 produces a basic diagnostic for picking the threshold. Above u the mean exceedance should be linear in u at points for which the model holds. It is proposed when looking at the empirical mean residual life plot to check for linearity. When selecting the threshold there is a bias-variance trade-off. If the threshold chosen is not big enough there is bias due to the model asymptotics being void and if the threshold chosen is too big the variance is big and this is because there is not a lot of data points (Coles and Davison, 2008). The next section, Section 2.3, will now go on to use the various techniques described in Section 2.2 in order to compare measured and modelled data.

2.3 Comparing Measured Data and ADMS-Urban Modelled Data

In this section, it is of interest to look at the output produced from both the monitoring site data, the diffusion tube data and the data produced by running the ADMS-Urban model to compare measured data (monitoring site and diffusion tube data) and modelled data (ADMS-Urban modelled data). All six monitoring sites will be looked at and the pollutant of interest is NO_2. As mentioned previously the year that is focused on is 2012. Average monthly NO_2 concentrations will be looked at to briefly examine some similarities between the measured data and the modelled data. Here the monitoring site, diffusion tube and ADMS-Urban modelled data will be considered. Then daily NO_2 concentrations will be investigated in greater detail for the monitoring site and modelled data to see how well both of these sets of data are calibrated.

Firstly, the monthly mean concentrations were calculated for the monitoring site, diffusion tube and ADMS-Urban modelled data and then the monthly maximum concentrations were calculated and this was only done for the monitoring site and ADMS-Urban modelled data. Both the observed data and the modelled data were plotted on the same time series plot so that it was easier to compare them. To identify the ADMS-Urban pixel closest to the monitoring site and diffusion tube locations, Euclidean distance was used. Each of the ADMS-Urban pixels were taken and the euclidean distance to the monitoring site or diffusion tube was calculated. The pixel that yielded the smallest distance was considered as the closest pixel to the location and that pixel was used for comparison.

Individual plots were produced for all six monitoring sites and since there were data available for 40 diffusion tubes in total these were all plotted on the one plot for convenience. The same procedure was done for the daily NO_2 concentrations but this time only monitoring site and modelled data were of interest. A plot of the daily mean differences between the days for both the modelled and monitoring data was also produced. Daily maximum concentrations were also calculated for the monitoring and modelled data and again a plot of the daily maximum differences between the days for both the modelled and monitoring data has been computed. These plots were again produced for all six monitoring sites.

Missing data occurred in three of the monitoring sites out of six for the daily NO_2 concentrations, namely Union Street Roadside, Errol Place and King Street. The percentage of missing data at these three sites were extremely low and these are presented in Table 2.2 below:

Table 2.2: Monitoring sites and the % of missing data

Monitoring Site	% of missing data
Union Street Roadside	1.37%
Errol Place	6.01%
King Street	7.65%

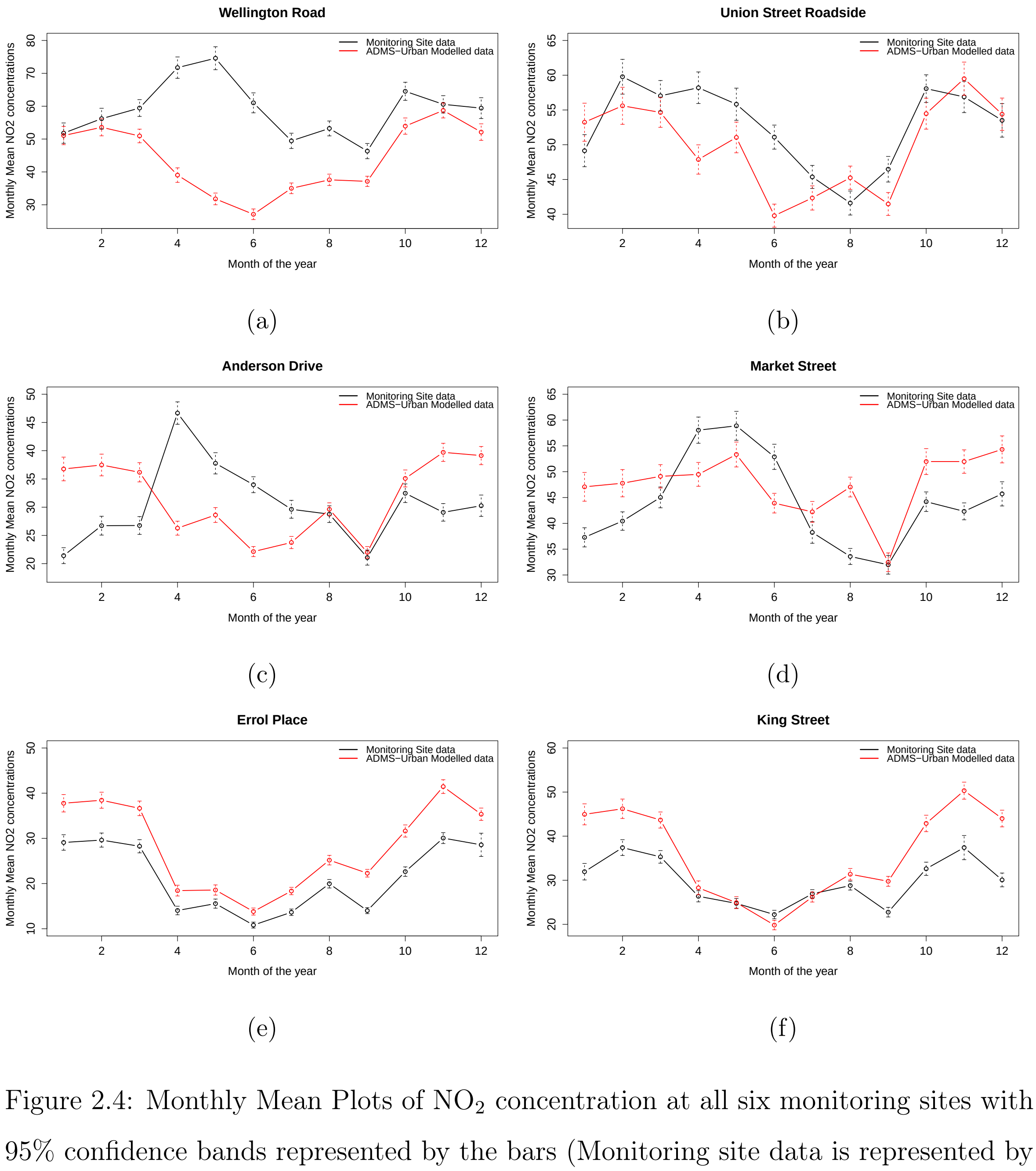
Missing data also occurred in thirteen of the diffusion tubes out of forty. For the analysis carried out in this chapter, missing data will not be an issue, however, it will become a difficulty later in the thesis.

To formally determine how well measured and modelled data are calibrated, errors in variables regression was carried out where the ADMS-Urban model were the response variable and the explanatory variable were the monitoring site data. Here the six sites were considered and the slope estimate was of concern for each site. Plots of the differences between the modelled and monitoring data were also produced for each site and it was of particular interest to see how close these values were to zero. Bland Altman plots (daily mean differences against daily mean averages) were also produced for all six sites. These methods have been carried out as they will give a better idea of how well calibrated the ADMS-Urban model and monitoring data are. Errors in variables regression was once again carried out but this time the ADMS-Urban model output was the response variable and the diffusion tube data were the explanatory variable. Extreme value analysis will be carried out and the aim here is to see if extreme values over particular thresholds occur at the same point in time. Additionally, the intensity values will be determined to examine the rates of exceedances. The extreme value analysis will only focus on the monitoring data for the six monitoring sites and the modelled data.

2.3.1 Monthly timescale

Monitoring Sites

In this section, the relationship between the monthly mean NO_2 concentrations for the monitoring and modelled data will be explored. The monthly mean maximum NO_2 concentrations will also be examined to investigate in terms of the peaks how well the modelled and monitoring data are calibrated. Firstly the monthly mean plots will be given followed by the monthly maximum plots.



(a) (b) (c) (d) (e) (f)

Figure 2.4: Monthly Mean Plots of NO_2 concentration at all six monitoring sites with 95% confidence bands represented by the bars (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

Figure 2.4a, highlights that at Wellington Road the model appears to under estimate

NO_2 concentrations throughout the full year of 2012 especially for the months of April through to June. Throughout these months the monthly profile shows that the model and monitoring data differ in terms of the average monthly NO_2 concentrations. The monitoring data increases then decreases which is in contrast to the model data which decreases then increases. From July onwards both sets of data appear to improve in terms of similarity.

Figure 2.4b shows that at Union Street Roadside in January the modelled data is slightly higher than the monitoring data and then is lower than the monitoring data from February until July. The modelled data is then higher than the monitoring data at Union Street Roadside for the month of August and then is lower again for September and October and for the latter part of the year it appears to be higher again. Looking at the monitoring data at Union Street Roadside it appears from the months February to August that there is general decrease in monthly mean NO_2 concentrations, then the monthly mean concentrations appear to increase again until October where they gradually decrease again. The model data appears to be a lot more varied over the year.

At Anderson Drive and Market Street it can be viewed from Figures 2.4c and 2.4d that both the model and monitoring monthly mean concentrations are varied over the year of 2012. At Anderson Drive the modelled data is higher than the observations measured by the monitoring station at the start of the year and again from August onwards. Meanwhile, the modelled data are lower than the monitoring data from the month of April to July. At Market Street the modelled data are higher than the monitoring data at the start of the year and from July onwards, however, they are lower from April to June.

The monthly profile clearly shows that at Errol Place even though the modelled data is higher than the monitoring data throughout the whole year of 2012, the pattern they follow is extremely similar. This highlights that the modelled data are slightly higher than the concentrations observed but they increase and decrease at the same rate. However, as previously mentioned in Section 2.1, for background concentrations, observed data from Errol Place was used for the baseline conditions so we would expect this level of agreement. At King Street, the modelled monthly mean NO_2 concentrations are higher than the monitoring monthly mean NO_2 concentrations for most of

2012 with the modelled data being slightly lower for June and July. At King Street the patterns of the monitoring and model data appear to be very alike. Figure 2.4 also emphasises the 95% confidence intervals of each monthly mean value at each monitoring site and these are represented by the bands around each of the data points. The monthly maximum concentrations were also calculated to investigate whether the ADMS-Urban model was performing well in terms of picking up the peaks of NO2 concentration.

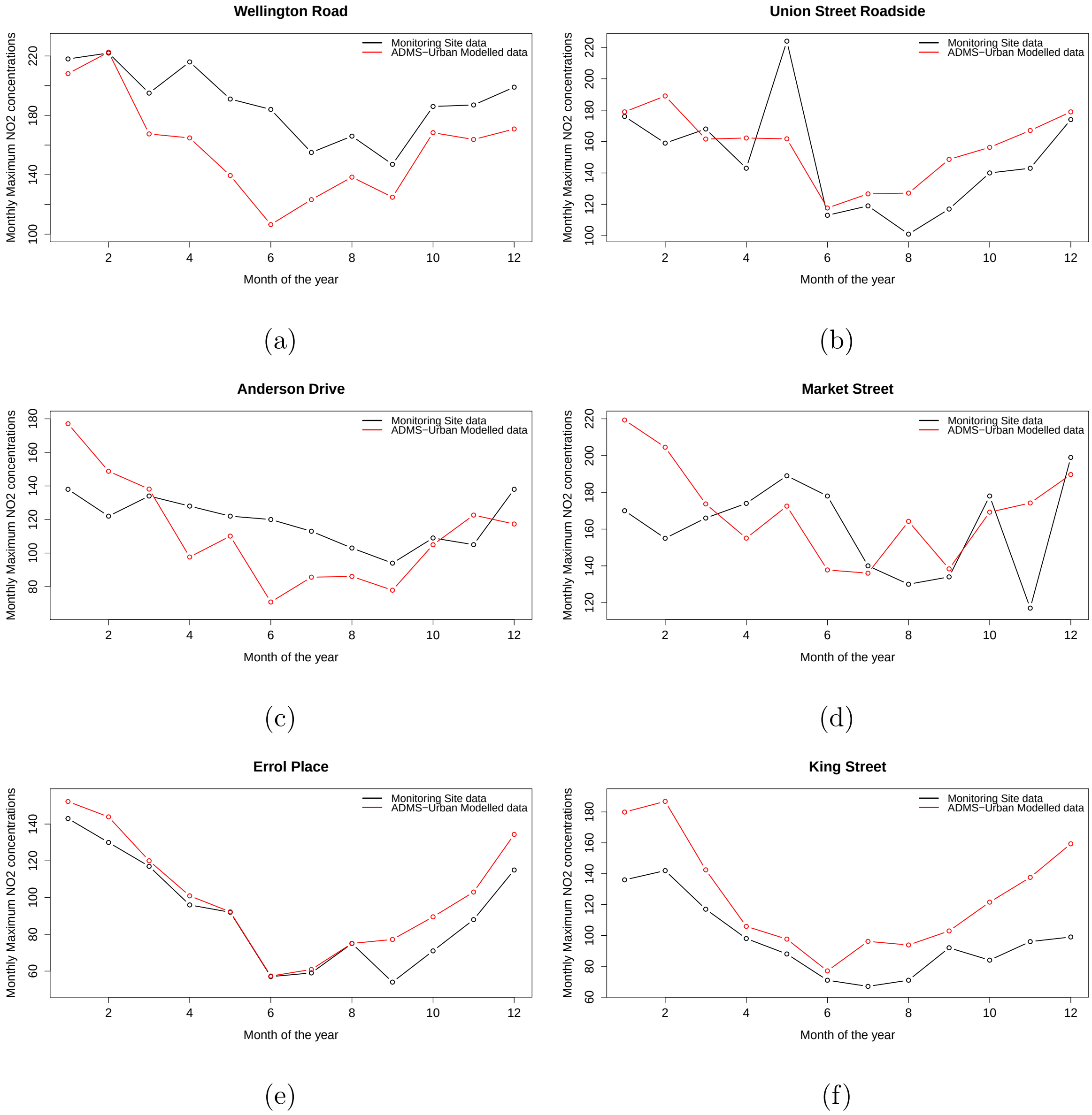


Figure 2.5: Monthly Maximum Plots of NO_2 concentration at all six monitoring sites (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

Figure 2.5a, highlights at Wellington Road that the ADMS-Urban modelled monthly maximum NO_2 concentrations are lower than the monitoring monthly maximum NO_2

concentrations throughout the timescale except for the month of February. The monthly maximum concentrations seem to more accurately follow the same pattern than the monthly mean concentrations at Wellington Road. Figure 2.5b highlights at Union Street Roadside that the monitoring data is slightly more varied than the model data especially for the first half of the year. From June onwards the model and monitoring data follow each other better in terms of pattern were the modelled data are higher than the monitoring data at Union Street Roadside.

From Figure 2.5c, it can be pointed out that at Anderson Drive for the year 2012 the monthly maximum concentrations produced by the model have more variability than those observed by the monitoring station. It can also be noted at the beginning of the year and for the month of November the modelled data is higher than the monitoring data and the rest of the year the modelled monthly maximum concentrations appear to be lower than the monitoring monthly maximum concentrations.

Furthermore, it is highlighted from Figure 2.5d that at Market Street the model and monitoring data both vary over the year. It can also be observed that the modelled data have a lot of variability throughout the year and change from being higher than the monitoring data to being lower than the monitoring data. Figure 2.5e shows that at Errol Place the modelled monthly maximum concentrations are higher than the monitoring monthly maximum concentrations throughout most of the year with both sets of data almost exactly the same for the months May to August. The pattern both sets if data follow at Errol Place is extremely similar which as mentioned previously, is to be expected. Figure 2.5f shows at King Street the modelled data is higher than the monitoring data for the whole of 2012. Again just like previously at King Street the pattern of the monitoring and model data appear to be very alike.

Diffusion Tubes

Various elements contribute to the precision of the diffusion tubes and the precision can differ depending on these elements. Consequently diffusion tubes are calibrated utilising a bias-adjustment factor achieved from co-location investigations (Pannullo *et al.* 2015). As well as monitoring data, looking at the comparison of the diffusion tube data and the modelled data was also of interest as the diffusion tube data are also observed in Aberdeen and cover a larger spatial region. Since there are 40 diffusion tube locations, showing a plot for every location would not be practical so instead the mean of all diffusion tube data was shown on the one plot, this gives 480 time points as $40 \times 12 = 480$. Data is only observed for monthly mean NO_2 concentrations as diffusion tubes record an integrated measure.

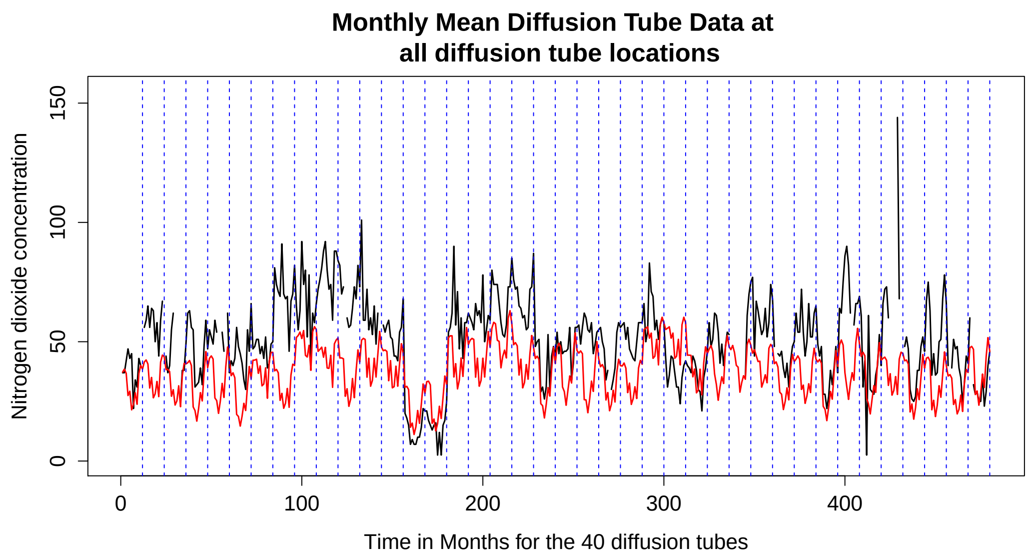
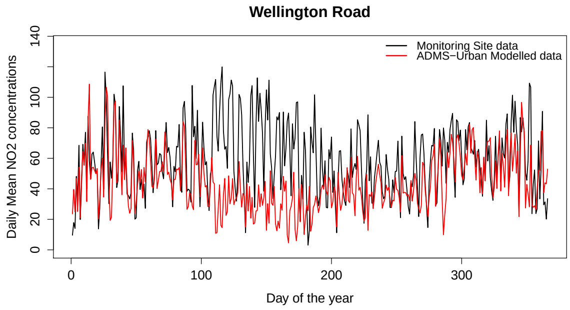


Figure 2.6: Monthly Mean Plot of NO_2 concentrations for diffusion tubes (Diffusion tube data is represented by the black line and the ADMS-Urban modelled data is represented by the red line). Vertical lines separate the data for each diffusion tube over the year.

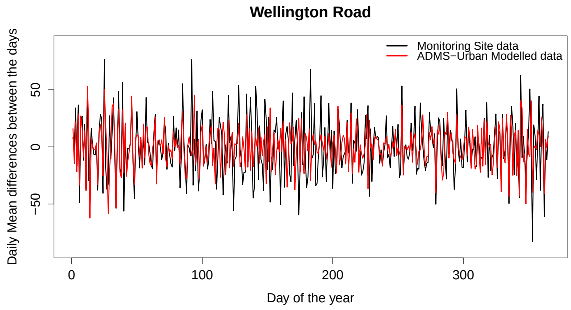
Figure 2.6 highlights that for the majority of time the ADMS-Urban model appears lower than the diffusion tube NO_2 concentrations. However, the level of concentrations reported seems to be fairly similar and both time series appear to follow each other in terms of pattern. It is also highlighted that the diffusion tube data appears to be more variable throughout Time.

2.3.2 Daily timescale

This section focuses on a higher temporal resolution, daily mean and maximum NO_2 monitoring and modelled concentrations are now compared. Sites are compared individually and daily mean plots are presented first followed by daily maximum plots.



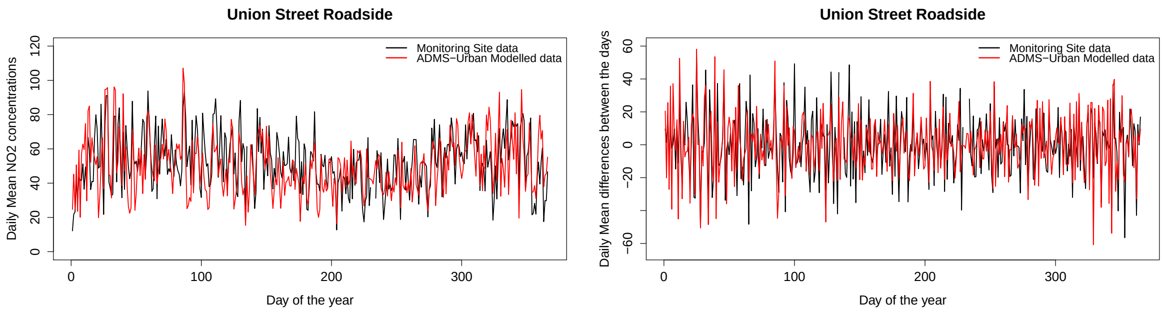
(a) Daily Mean NO_2 data



(b) Daily Mean difference between the days

Figure 2.7: Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Wellington Road (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

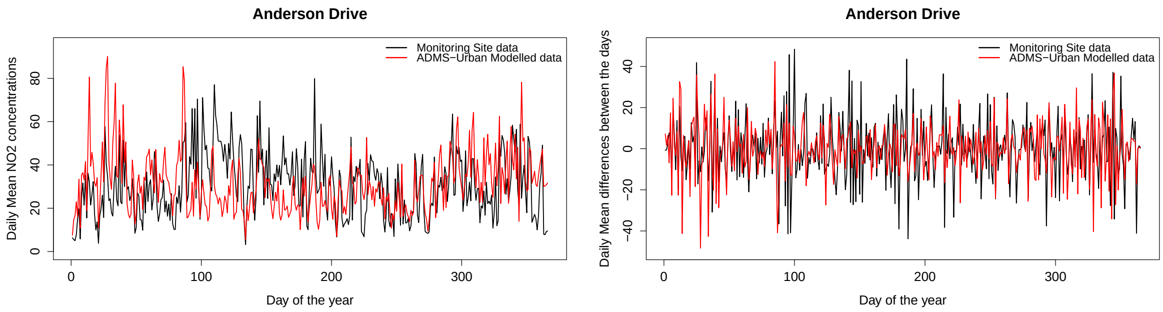
Looking at the daily profile in Figure 2.7a, highlights that the modelled data appears to be lower than daily mean observed NO_2 concentrations from around day 100 (9^{th} April) of the year to around day 200 (18^{th} July) of the year. Following that, the model and monitoring sites profiles are much closer, although there still seems to be time points where ADMS-Urban modelled data is lower throughout the year. The monitoring site data also seem much more varied over the short timescale. Figure 2.7b expresses the variability in the differences of the monitoring and modelled data day to day.



(a) Daily Mean NO_2 data (b) Daily Mean difference between the days

Figure 2.8: Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Union Street Roadside (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

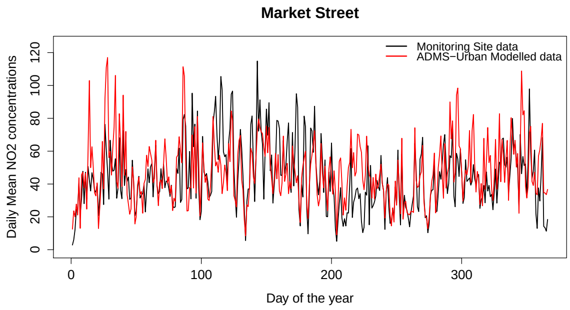
From Figure 2.8a it can be seen, at Union Street Roadside, that even though the modelled data appears to be slightly lower than the daily mean observed NO_2 concentrations from around day 100 of the year to around day 200 of the year, the model and monitoring data do seem to be well calibrated. It also appears as though the model data has more variability than the monitoring data. Figure 2.8b highlights the day to day variation in the differences and suggests that day to day the differences in the modelled data appear to be larger.



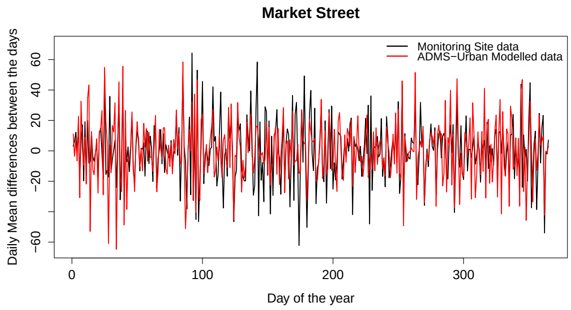
(a) Daily Mean NO_2 data (b) Daily Mean difference between the days

Figure 2.9: Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Anderson Drive (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

Figure 2.9a, highlights at Anderson Drive, that the modelled data seems to be higher than the observed data at the start of year until around day 100, and then it seems to be lower from day 100 to around day 200 just as it does at Wellington Road and Union Street Roadside. Subsequently, at the latter part of the year the model appears to over predict again. Figure 2.9b suggests that the modelled data daily differences day to day are more varied at the beginning of the year and the end of the year and the observed differences day to day are more varied around the middle of the year from about day 100 to day 200.



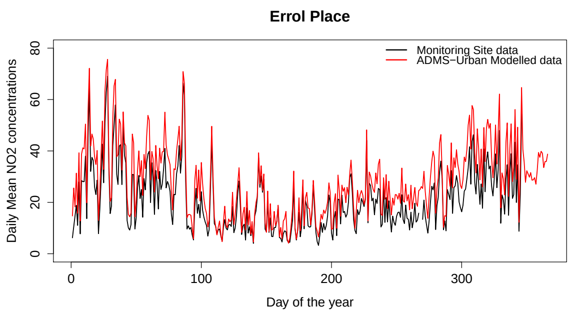
(a) Daily Mean NO_2 data



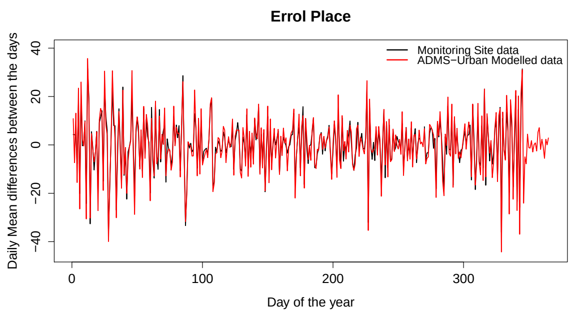
(b) Daily Mean difference between the days

Figure 2.10: Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Market Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

Looking at Figure 2.10, the model and monitoring data seem to behave similarly at Market Street to Anderson Drive with the same patterns in the data occurring. This can be observed in both Figures 2.10a and 2.10b.



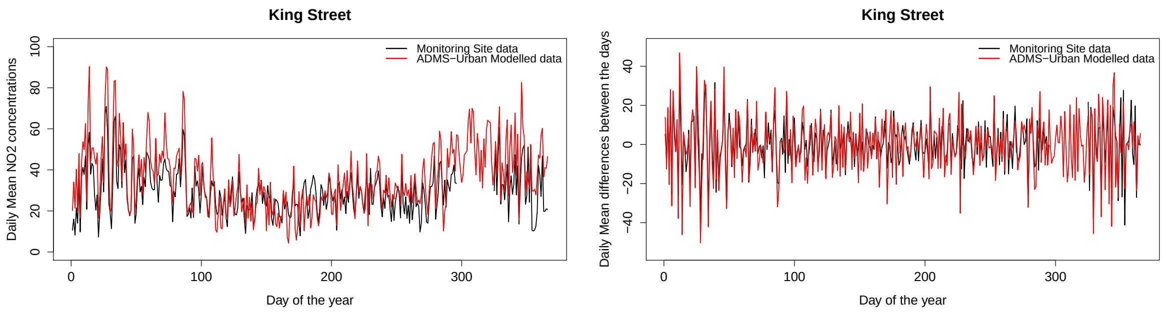
(a) Daily Mean NO_2 data



(b) Daily Mean difference between the days

Figure 2.11: Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Errol Place (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

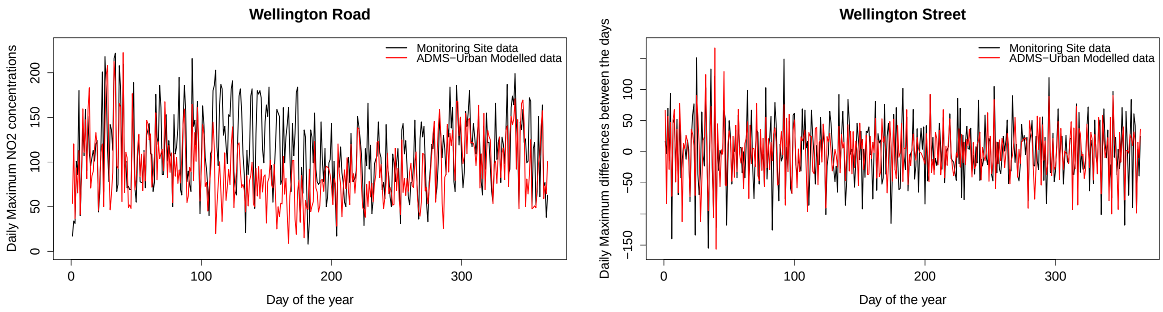
Figure 2.11a highlights that at Errol Place the modelled data appears to be slightly higher than the daily mean observed NO_2 concentrations throughout the year of 2012. The modelled data also seem much more varied over the short timescale. The daily mean differences emphasises that day to day concentrations produced by the model are a lot more varied as the differences are considerably greater. This appears to be the case throughout the full year with it being more visible at the beginning and end of the year.



(a) Daily Mean NO_2 data (b) Daily Mean difference between the days

Figure 2.12: Daily Mean and Daily Mean Difference between Days Time Series Plots of NO_2 concentration at the monitoring site King Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

For King Street, Figure 2.12b highlights that the day to day differences in the modelled data appear to have more variability than the day to day differences in the monitoring observations. Figure 2.12a emphasises that the modelled data seems to be higher than the monitoring data at some points in time over the year at King Street. A similar comparison of the daily maximum concentrations was also carried out to examine whether the ADMS-Urban model was performing well in terms of picking up the peaks of NO_2 concentration. This will only be shown for Wellington Road and Union Street Roadside and the rest of the plots may be seen in Appendix A.

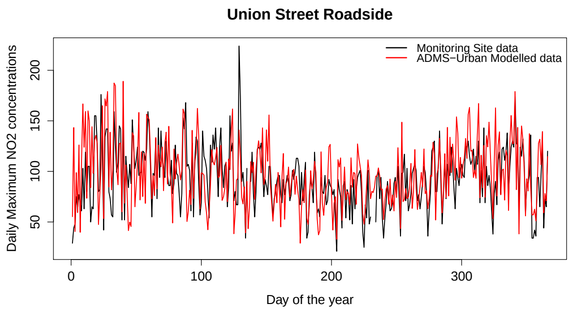


(a) Daily Maximum NO_2 data

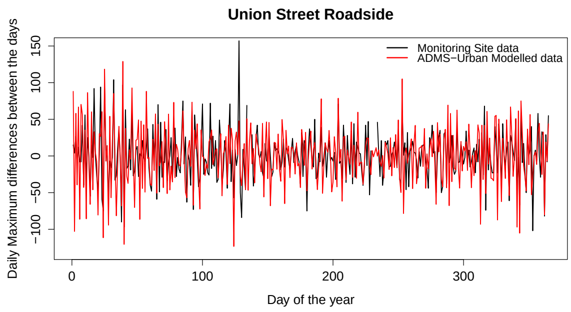
(b) Daily Maximum difference between the days

Figure 2.13: Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Wellington Road (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

Looking at Figure 2.13a highlights, that just like the daily mean NO_2 concentrations, the modelled data appears to be lower than the daily maximum observed NO_2 concentrations from around day 100 of the year to around day 200 of the year. Just as before from Figure 2.7a it can be highlighted from Figure 2.13a that the model and monitoring sites profiles are much closer after day 200. However, there still appears to be data produced from the model that are slightly lower compared with the monitoring data after this time point and this continues throughout the year 2012. Both the model and monitoring data appear to be varied over the year. Figure 2.13b highlights that both the modelled and monitoring data are varied in terms of differences in day to day NO_2 concentrations.



(a) Daily Maximum NO_2 data



(b) Daily Maximum difference between the days

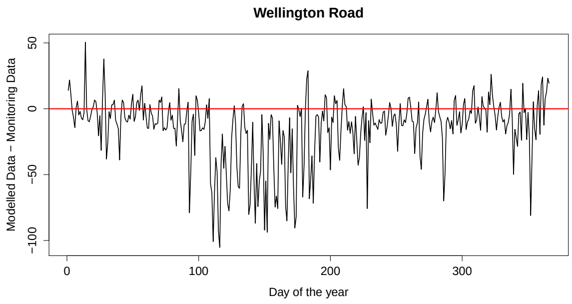
Figure 2.14: Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Union Street Roadside (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

From Figure 2.14a it can be seen, at Union Street Roadside, that the model and monitoring data appear to be well calibrated. It can be clearly seen that the modelled data are much more varied throughout the year of 2012. Figure 2.14b suggests that the daily maximum day to day differences are much more varied over the year for the modelled data.

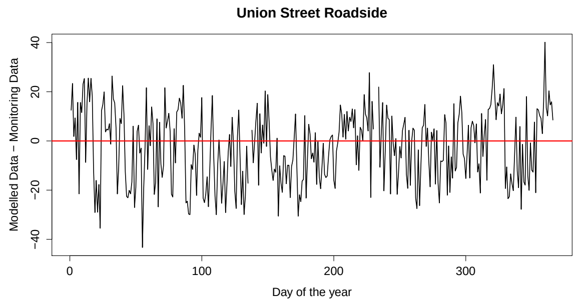
Now formal analysis is carried out in order to determine how well the ADMS-Urban modelled data and measured data are calibrated. All formal analysis were carried out on the daily mean NO_2 concentrations. However, errors in variables regression were also carried out with the monthly NO_2 modelled data as the response variable and the monthly NO_2 diffusion tube data as the explanatory variable. This was done in order to see formally how well the model is related to the diffusion tube observations. It was mentioned previously that as well as errors in variables regression, difference plots (Modelled data - Monitoring data), bland altman plots and extreme value analysis were carried out. Firstly the difference plots will be considered followed by the bland altman plots and then errors in variables regression and extreme value analysis results will be produced.

2.4 Formal Assesement of Comparing Measured data and Modelled Data on a Daily timescale

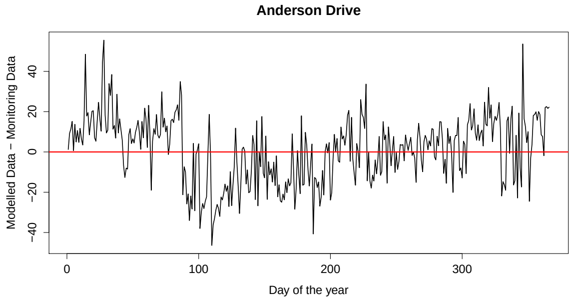
2.4.1 Differences



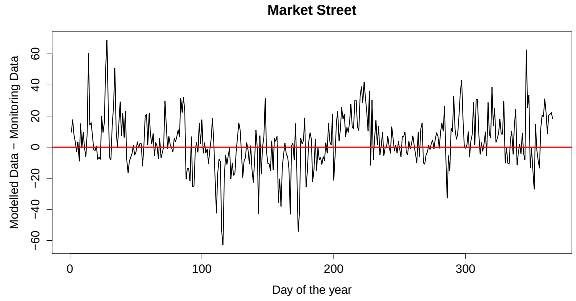
(a)



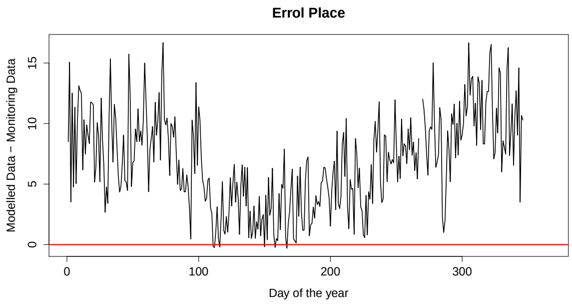
(b)



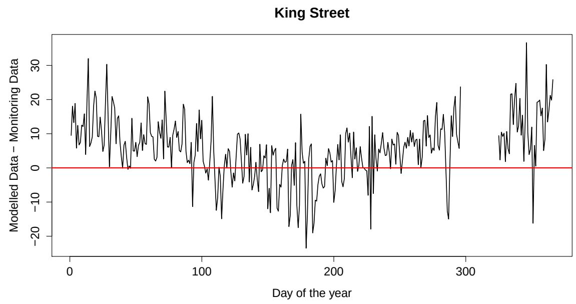
(c)



(d)

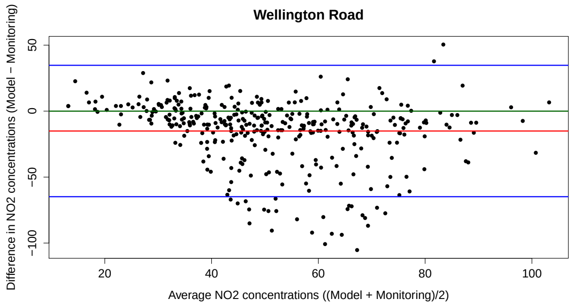


(e)

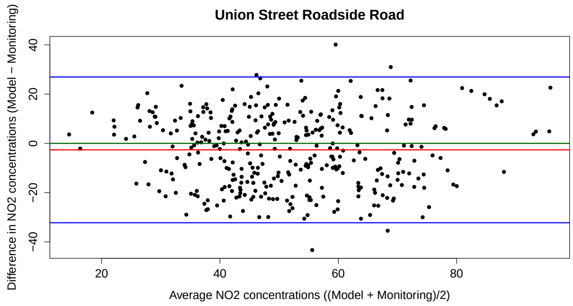


(f)

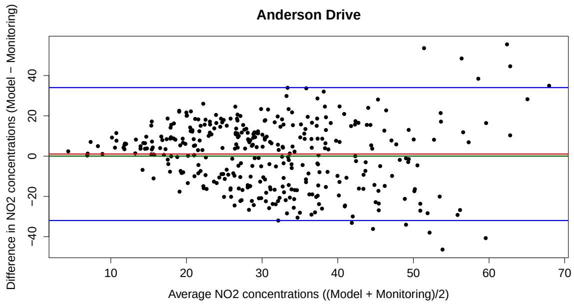
Figure 2.15: Daily Mean Differences of the modelled data (μgm^{-3}) minus the monitoring data (μgm^{-3}) where the red line represents the horizontal line, $x = 0$



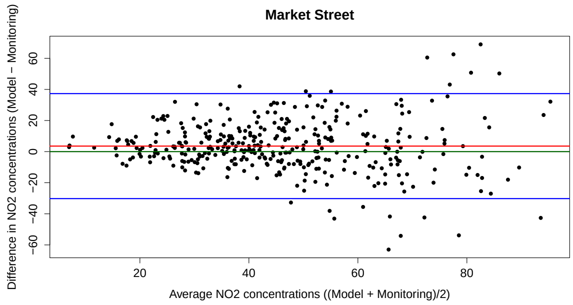
(a)



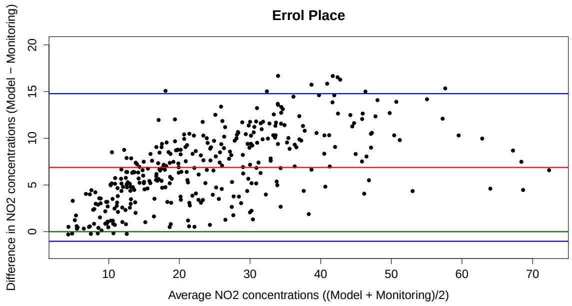
(b)



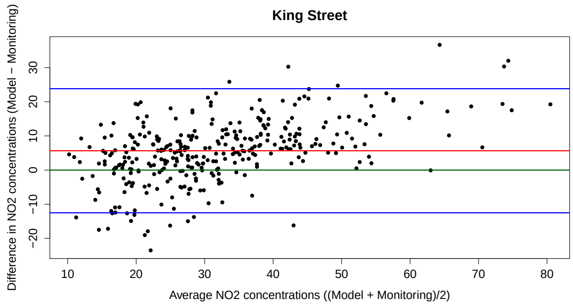
(c)



(d)



(e)



(f)

Figure 2.16: Bland Altman plots (Daily Mean Differences against Daily Mean Averages) where the red line represents the mean difference, the blue lines represent the lower and upper 95% confidence intervals and the green line represents the horizontal line, $x = 0$

The first formal assessment of comparing measured data and modelled data on a daily timescale was to examine the daily mean differences and the daily mean differences against the daily mean averages. The differences were calculated by taking the monitoring data away from the modelled data. If in Figure 2.15 the data are centred around zero then this would mean that the modelled and monitoring data were giving the exact same value and that the modelled and monitoring data are both extremely well calibrated. From Figure 2.15a it can be seen that the monitoring site Wellington Road appears to have the highest difference occurring with some as large as 100 μgm^{-3} where the modelled data appears to be lower than the monitoring data. This can be seen to occur at around the same time that the modelled data appeared to be lower than the monitoring data in the daily mean profile at Wellington Road. The predictions at Errol Place, observed in Figure 2.15e, appear to be the best of the six sites with the modelled data higher than the monitoring data by as much as 15 μgm^{-3} which would be expected as it has been previously stated that Errol Place data were incorporated into the model run. At Anderson Drive and Union Street Roadside the modelled data appears to be lower and higher than the monitoring data by as much as 40 μgm^{-3} and this increases to 60 μgm^{-3} at Market Street. From Figure 2.15f and, the modelled data at King Street appear to be well calibrated with the modelled data higher than the monitoring data by 30 μgm^{-3} and lower than the monitoring data by 20 μgm^{-3}. This highlights that the modelled and monitoring data appears to be better calibrated at King Street and not as well at Market Street and Wellington Road. These differences reinforce what was seen in the daily mean profiles for all six of these sites. Figure 2.16 suggests that as the variation in differences increases, the average increases. This can be observed mainly in Figures 2.16a, 2.16c and 2.16d and this emphasises that bias is not consistent and depends on the level of measurement. In Figures 2.16e and 2.16f there appears to be a linear relationship between the differences and the average of the modelled and monitoring data. There appears to be no relationship between the differences and the average of the modelled and monitoring data in Figure 2.16b and the points appear to be roughly equally scattered above and below the zero line. A table of the mean and variance of these differences and 95% confidence intervals for the mean difference is given in Table 2.3.

Table 2.3: Mean difference, Variance difference and 95% Confidence Intervals (CI) for the mean difference

Site	Mean	Variance	95% CI for the mean difference
Wellington Road	-15.067	619.894	(-17.618, -12.516)
Union Street Roadside	-2.635	218.967	(-4.162, -1.109)
Anderson Drive	1.031	272.229	(-0.659, 2.721)
Market Street	3.551	284.943	(1.822, 5.281)
Errol Place	6.876	15.606	(6.459, 7.294)
King Street	5.654	82.478	(4.686, 6.622)

According to the information on Table 2.3 it can be suggested that on average at

Wellington Road and Union Street Roadside, the modelled data is lower than the observed NO_2 concentrations. Meanwhile, at the other monitoring sites on average the ADMS-Urban modelled data is higher than the observed NO_2 concentrations. Furthermore at Anderson Drive it can be observed from Table 2.3 that the 95% CI for the mean difference includes the value 0 suggesting at Anderson Drive that the modelled and monitoring data are not different. Overall, the variability in these differences appears to be extremely high except at Errol Place and King Street and especially for Wellington Road. This highlights that the modelled data and monitoring data appear to be more similar at Errol Place as would be expected and King Street and dissimilar at Wellington Road.

2.4.2 Deming Regression Results

Monitoring Data

For all deming regression analysis carried out throughout this chapter, the *MethComp* package (CRAN, 2013a) in *R* was used. After running both the deming regression and linear regression where the ADMS-Urban modelled data were the response variable and the monitoring site data were the explanatory variable, the following results were produced:

Table 2.4: Summary of Linear Regression Model

Site	Slope Estimate	Standard Error Estimate	95% Confidence Interval for the Slope
Wellington Road	0.2529	0.0391	(0.1760, 0.3298)
Union Street Roadside	0.5993	0.0418	(0.5172, 0.6814)
Anderson Drive	0.3089	0.0455	(0.2194, 0.3984)
Market Street	0.6072	0.0392	(0.5301, 0.6842)
Errol Place	1.1607	0.0154	(1.1303, 1.1910)
King Street	1.1542	0.0430	(1.0697, 1.2387)

Table 2.5: Summary of Deming Regression Model

Site	Slope Estimate	Standard Error Estimate	95% Confidence Interval for the Slope
Wellington Road	0.4998	0.0864	(0.3363, 0.6730)
Union Street Roadside	0.9876	0.0661	(0.8601, 1.1249)
Anderson Drive	0.7862	0.1533	(0.5296, 1.1220)
Market Street	0.9416	0.0780	(0.8012, 1.1068)
Errol Place	1.2015	0.0218	(1.1621, 1.2452)
King Street	1.4943	0.0591	(1.3886, 1.6209)

When running the deming regression the ratio of the error variances were assumed to be 1. Linear regression was run in order to show that there is errors in both sets of measurements. From Tables 2.4 and 2.5 it can be seen that the slopes produced from linear regression all appear to be smaller. This indicates that if errors in both sets of measurements are not considered then the slope is underestimated and thus highlights the importance of running deming regression. Table 2.5 emphasises at Wellington Road for every 1 μgm^{-3} increase in the monitoring data, on average the modelled data increases by 0.4998 μgm^{-3}. Looking at the 95% confidence intervals for the slope in Table 2.5 it could also be concluded at Wellington Road for every 1 μgm^{-3} increase in the monitoring data, on average the modelled data increases by somewhere between 0.3363 μgm^{-3} and 0.6730 μgm^{-3}. The other five monitoring sites can be explained in a similar manner. This further suggests that the modelled and monitoring data appear not to be well calibrated at Wellington Road and at the other monitoring sites they appear to be better calibrated.

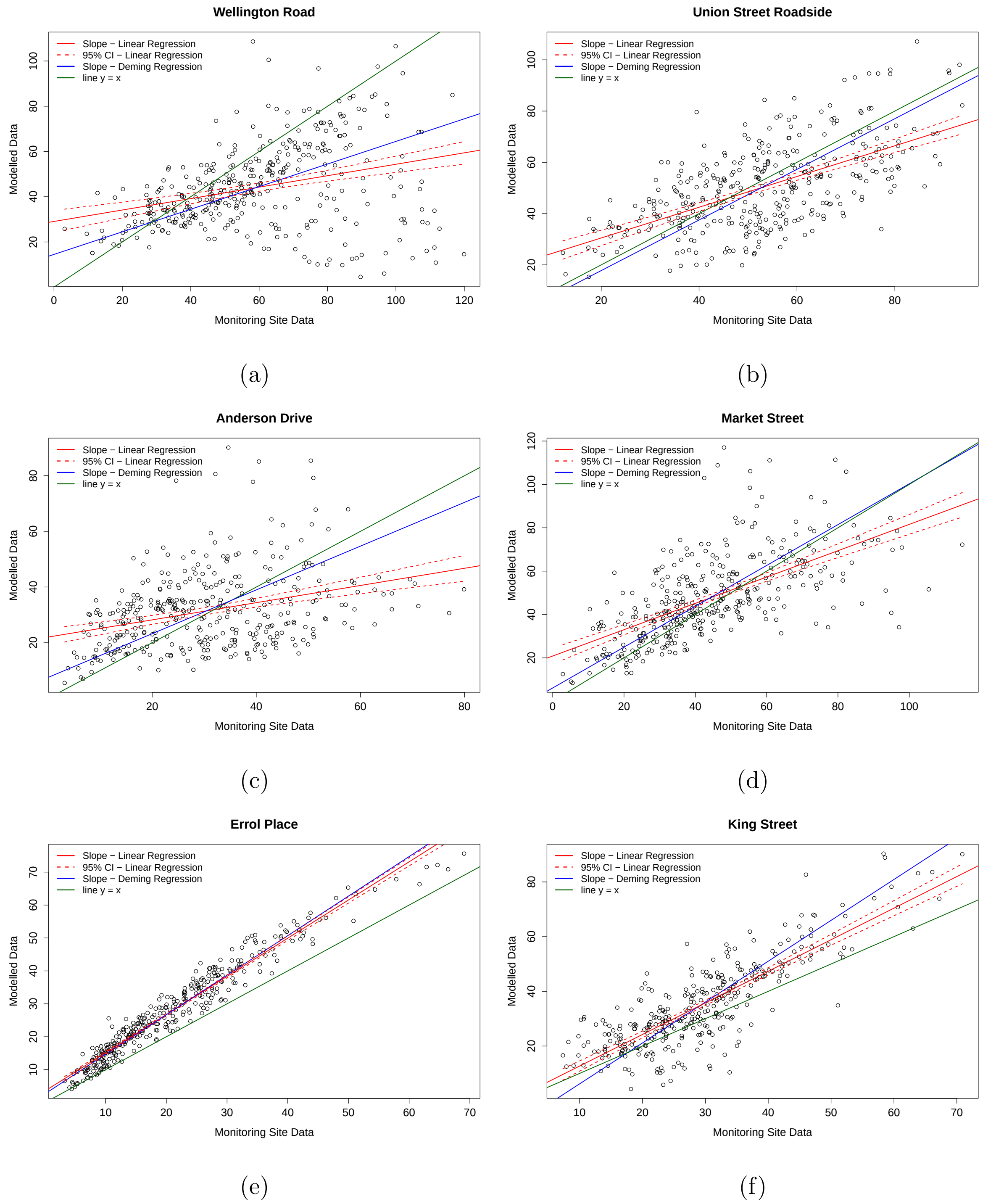


Figure 2.17: Scatterplots of Modelled data (μgm^{-3}) against Monitoring data (μgm^{-3}) with linear and deming regression lines, 95% confidence intervals for the linear regression line and the line $y = x$

Figure 2.17 highlights scatterplots of the modelled data against the monitoring data with both the deming and linear regression lines, 95% confidence bands for the linear regression line and the line $y = x$. Figure 2.17 highlights that the deming regression line appears to be lower than the linear regression line for all sites at lower NO_2 concentrations. Then as the concentrations increase the deming regression line appears to be higher than the linear regression line. This further emphasises that if errors in both sets of measurements are not considered then the slope estimates may be estimated incorrectly.

Diffusion Tube Data

Table 2.6 represents the results produced when carrying out linear regression and deming regression where the ADMS-Urban modelled data were the response variable and the diffusion tube data were the explanatory variable.

Table 2.6: Summary of Linear Regression and Deming Regression Model

Regression carried out	Slope Estimate	Standard Error Estimate	95% Confidence Interval for the Slope
Linear Regression	0.2954	0.0241	(0.2480, 0.3429)
Deming Regression	0.3863	0.0411	(0.3061, 0.4662)

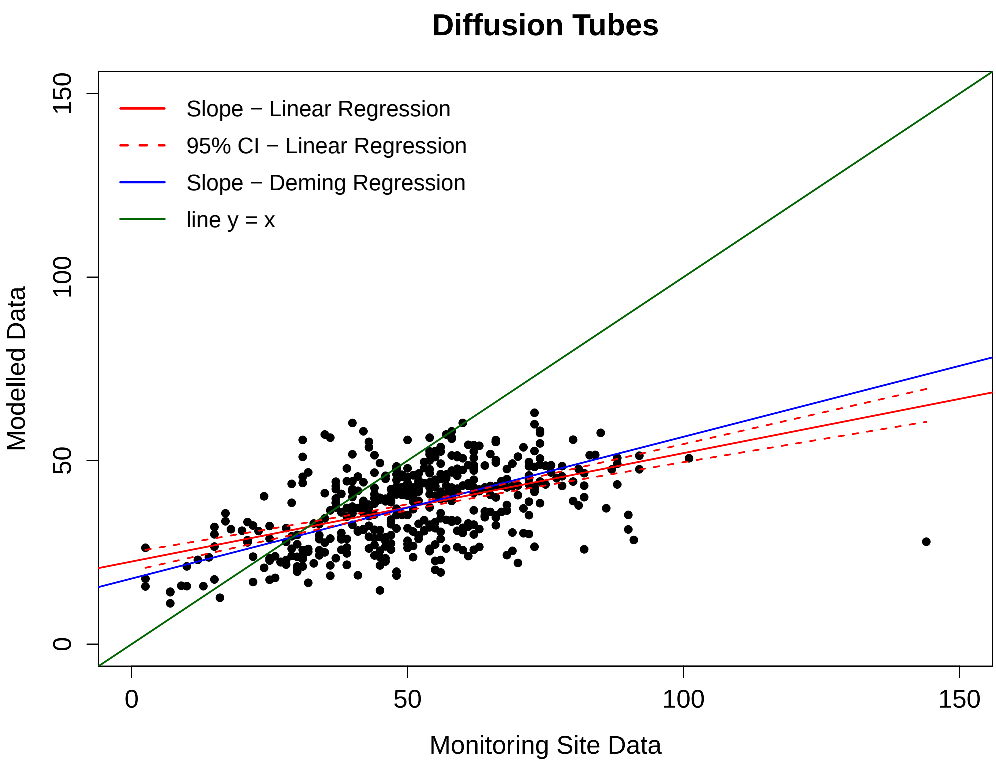


Figure 2.18: Scatterplot of Modelled data (μgm^{-3}) against Diffusion Tube data (μgm^{-3}) with linear and deming regression lines, 95% confidence intervals for the linear regression and the line $y = x$

Once again when running the deming regression the ratio of the error variances were assumed to be 1. From Table 2.6 it is highlighted once more that the slope produced from carrying out linear regression appears to be smaller. This further indicates that if errors in both sets of measurements are not considered then the slope is underestimated. It is emphasised from Table 2.6 that for every 1 μgm^{-3} increase in the diffusion tube data, on average the modelled data increases by 0.3863 μgm^{-3} when deming regression is considered. Looking at the 95% confidence interval for the slope in Table 2.6 it could also be concluded that for every 1 μgm^{-3} increase in the diffusion tube data, on average the modelled data increases by somewhere between 0.3061 μgm^{-3} and 0.4662 μgm^{-3} again when considering deming regression. This suggests that overall, the modelled data and diffusion tube data are not well calibrated at the diffusion tube locations. Figure 2.18 highlights scatterplots of the modelled data against the diffusion tube data with both the deming and linear regression lines, 95% confidence bands for the linear regression line and the line $y = x$. Figure 2.18 highlights that once more the deming regression line appears to be lower than the linear regression line at lower NO_2 concentrations. Then just like before for the monitoring sites, as the concentrations increase the deming regression line appears to be higher than the linear regression line. This further emphasises that if errors in both sets of measurements are not considered then the slope estimates may be estimated incorrectly.

2.4.3 Peaks over Threshold Results

The analysis so far has focussed mainly on the daily mean NO_2 concentrations but in an air quality context it may be the high values which are of interest so in this section the analysis focusses on these extremes. Exceedances above a threshold approach will be investigated to see if both sets of data (modelled and monitoring data) are behaving similarly i.e to see if the time points at which each set of data exceed the given threshold are roughly the same. Firstly, for this to be achieved a high threshold i.e the 90^{th} percentile will be investigated and this is just to explore if the modelled and monitoring peaks occur at roughly the same time. Then formally, thresholds will be chosen through mean residual life plots. Given these thresholds have been chosen correctly, the number of events exceeding these thresholds follow a Poisson distribution. Therefore, the number of events exceeding the given threshold can be investigated. For this analysis, the package *fExtremes* (CRAN, 2013b) in *R* was used.

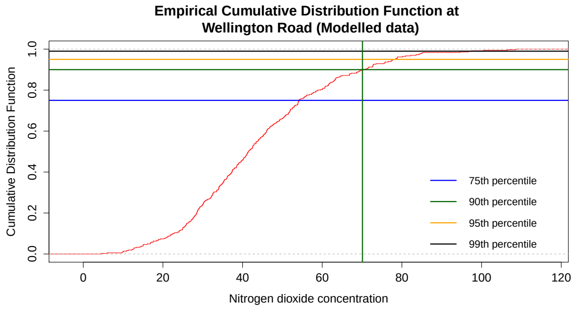
The 90^{th} percentile was chosen as it gave a sufficient number of exceedances to observe if the modelled and monitoring data were occurring at the same points in time over the year 2012. The 75^{th}, 95^{th} and 99^{th} percentiles were also investigated and the number of exceedances for each are given in Table 2.7 below.

Table 2.7: Number of exceedances over the 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles for both the modelled and monitoring data

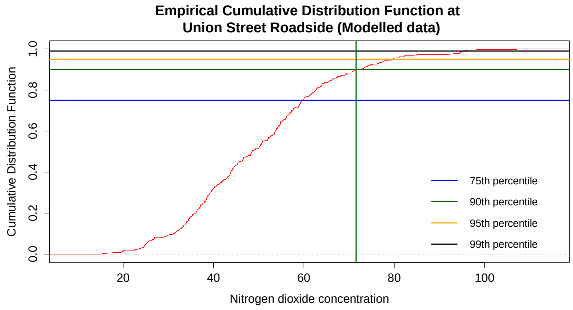
Percentile	Monitoring Site	Number of exceedances (Modelled data)	Number of exceedances (Monitored data)
75^{th}	Wellington Road	92	92
	Union Street Roadside	92	90
	Anderson Drive	92	92
	Market Street	92	91
	Errol Place	92	86
	King Street	92	85
90^{th}	Wellington Road	37	37
	Union Street Roadside	37	36
	Anderson Drive	37	37
	Market Street	37	37
	Errol Place	37	35
	King Street	37	34
95^{th}	Wellington Road	19	19
	Union Street Roadside	19	18
	Anderson Drive	19	19
	Market Street	19	19
	Errol Place	19	18
	King Street	19	17
99^{th}	Wellington Road	4	4
	Union Street Roadside	4	4
	Anderson Drive	4	4
	Market Street	4	4
	Errol Place	4	4
	King Street	4	4

Figures 2.19 and 2.20 represent the empirical Cumulative Distribution Functions (CDF) at all six monitoring sites for both the modelled and monitoring data respectively. From these plots it can be observed from the green vertical line the point at which the 90^{th} percentile cuts the CDF. These values are then used in Figure 2.21 to represent the

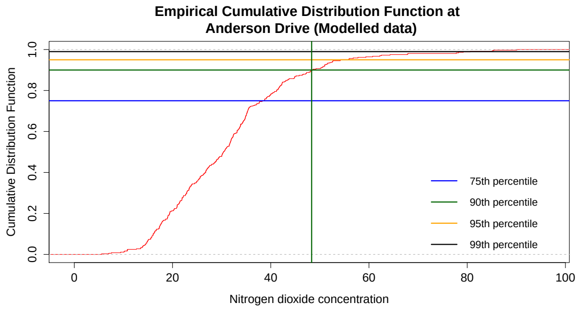
level at which these exceedances occur and to highlight how similar these values/levels are for both the modelled and monitoring data.



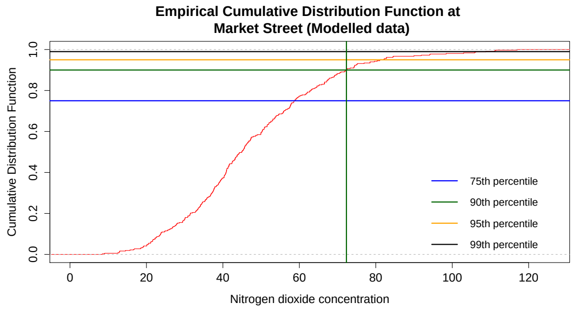
(a)



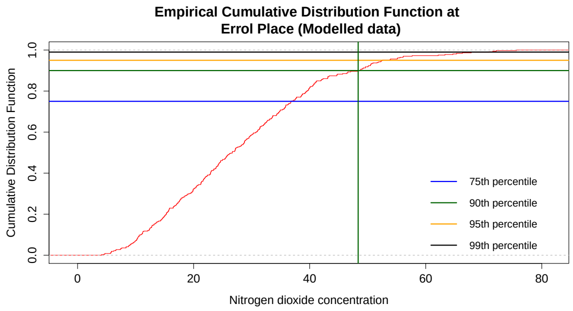
(b)



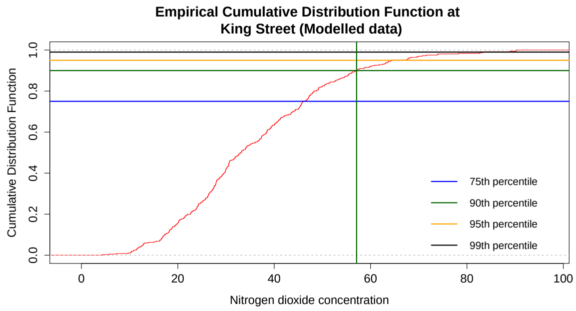
(c)



(d)



(e)



(f)

Figure 2.19: Empirical CDF at all six monitoring sites (Modelled data) with horizontal lines representing the 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles and the green vertical line represents the value at which the 90^{th} percentile cuts the CDF

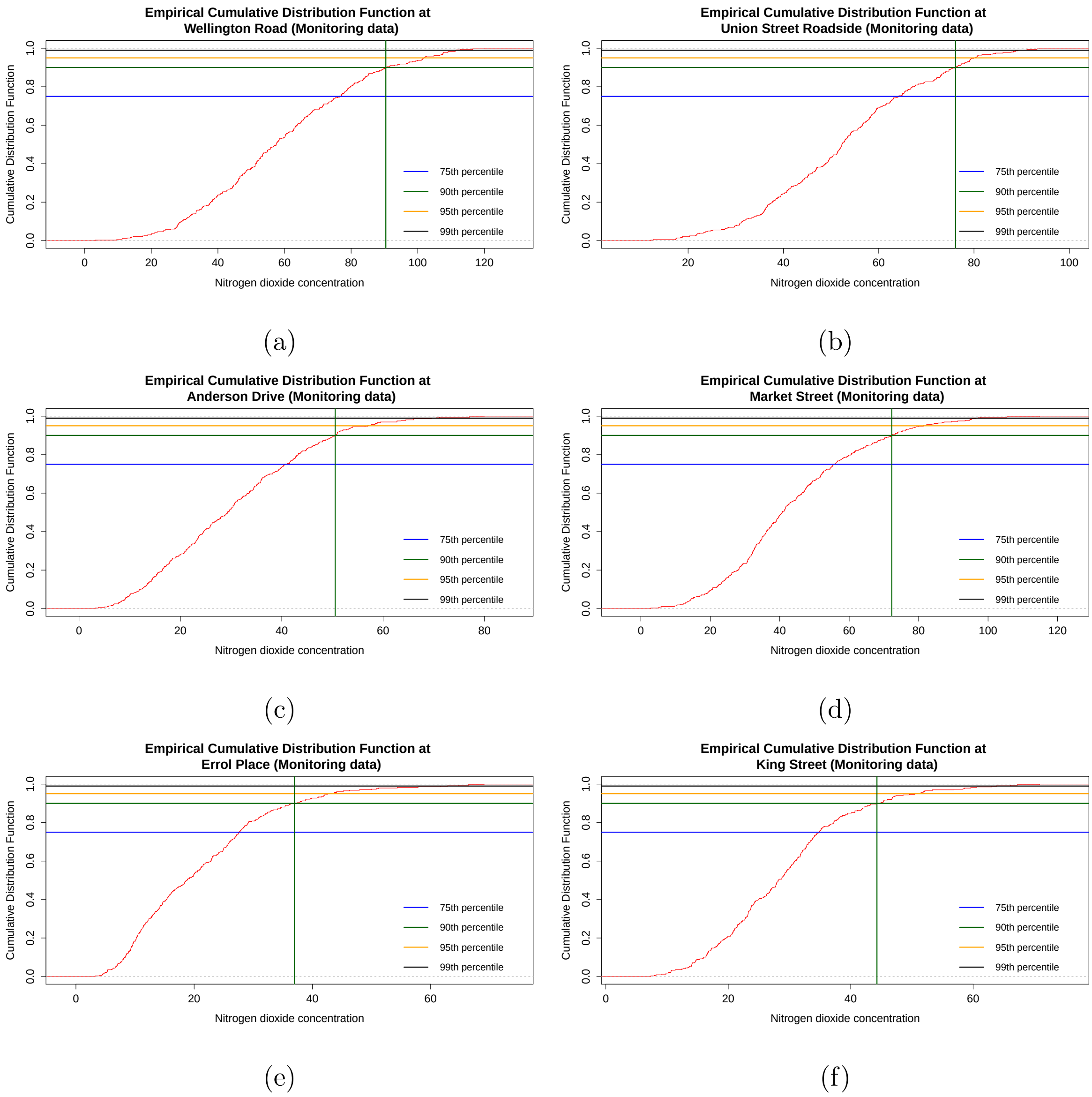


Figure 2.20: Empirical CDF at all six monitoring sites (Monitoring data) with horizontal lines representing the 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles and the green vertical line represents the value at which the 90^{th} percentile cuts the CDF

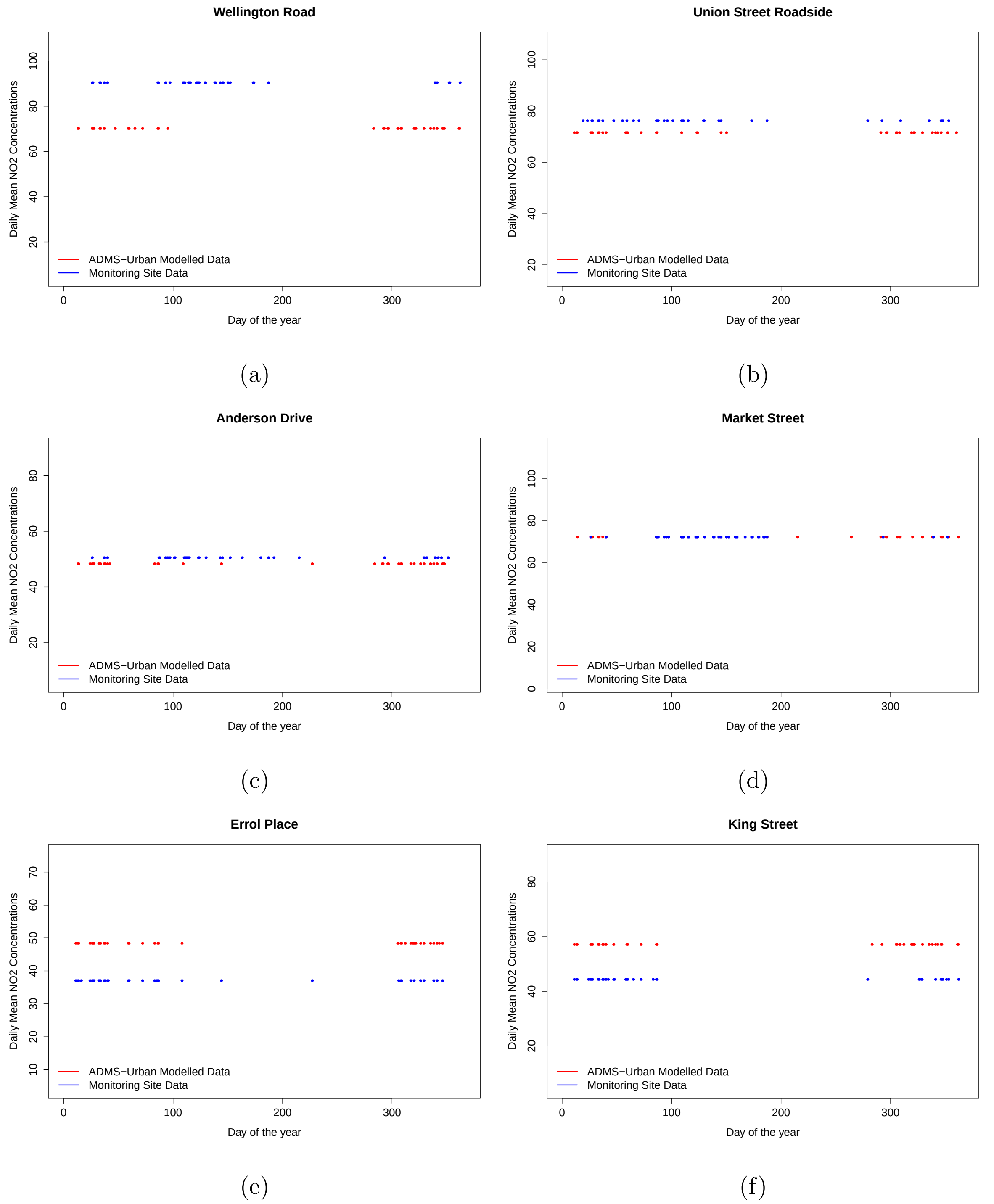


Figure 2.21: The time points at which the daily mean NO_2 modelled data and the daily mean NO_2 monitoring site data exceed the 90^{th} percentile

From Figure 2.21 it can be stated apart from the obvious comment about the different levels, at the monitoring site Errol Place and King Street, it is clear that both the model and monitoring daily mean NO_2 data are exceeding their threshold at the same points in time, with two extra points exceeding the threshold for the monitoring site data at Errol Place. However, at Wellington Road and Anderson Drive observed in Figures 2.21a and 2.21c, there are much less consistency in time of exceedance between the two series. It should be noted that the levels seem to be more similar for the monitoring sites Anderson Drive, Market Street and Union Street Roadside. From Figure 2.21b, at Union Street Roadside it appears again that the model and monitoring data are exceeding their threshold at the same points in time. Meanwhile, at Market Street there appears to be much less consistency in time of exceedance between the two series. Below Tables 2.8 and 2.9 have been produced to investigate whether the exceedances above the 90^{th} percentile have any dependence on the months of the year and if these dependences are similar for both sets of data.

Table 2.8: Number of Exceedances over the 90^{th} percentile for both the modelled and monitoring data for months Jan to June over the year 2012

Data	Site	Jan	Feb	Mar	Apr	May	June
Modelled data	Wellington Road	5	6	4	1	0	0
	Union Street Roadside	6	7	3	1	4	0
	Anderson Drive	6	7	3	1	1	0
	Market Street	4	4	2	3	6	0
	Errol Place	7	8	4	1	0	0
	King Street	6	8	3	0	0	0
Monitoring data	Wellington Road	2	4	2	9	12	2
	Union Street Roadside	4	6	5	7	4	1
	Anderson Drive	1	2	2	11	6	2
	Market Street	1	1	3	8	10	8
	Errol Place	8	9	5	1	1	0
	King Street	7	11	5	0	0	0

Table 2.9: Number of Exceedances over the 90^{th} percentile for both the modelled and monitoring data for months July to Dec over the year 2012

Data	Site	July	Aug	Sept	Oct	Nov	Dec
Modelled data	Wellington Road	0	0	0	6	8	7
	Union Street Roadside	0	0	0	4	6	6
	Anderson Drive	0	1	0	5	8	5
	Market Street	1	1	1	4	5	6
	Errol Place	0	0	0	1	12	4
	King Street	0	0	0	3	10	7
Monitoring data	Wellington Road	1	0	0	0	0	5
	Union Street Roadside	1	0	0	2	2	4
	Anderson Drive	2	1	0	1	3	6
	Market Street	3	0	0	1	0	2
	Errol Place	0	1	0	0	7	3
	King Street	0	0	0	1	3	7

Tables 2.8 and 2.9 further highlight that at the monitoring sites Errol Place and King Street, it is clear that both the model and monitoring daily mean NO_2 data are exceeding their threshold at roughly the same points in time. However, at Wellington Road and Anderson Drive and Market Street there are much less consistency in time of exceedance between the two series. Meanwhile, at Union Street Roadside the model and monitoring data are exceeding the 90^{th} percentile again at roughly the same points in time, although not as well as King Street and Errol Place. It can also be highlighted that exploring the number of exceedances for each month of the year that the compared with the summer, in the winter the monitoring and modelled data appear to have a similar amount of exceedances.

Mean Residual Life Plots have now been investigated to formally determine thresholds. The number of events exceeding these thresholds can be achieved given these thresholds have been chosen correctly as the number of events follow a Poisson distribution. Figures 2.22 and 2.23 below represent the mean residual life plots for both the model and monitoring data.

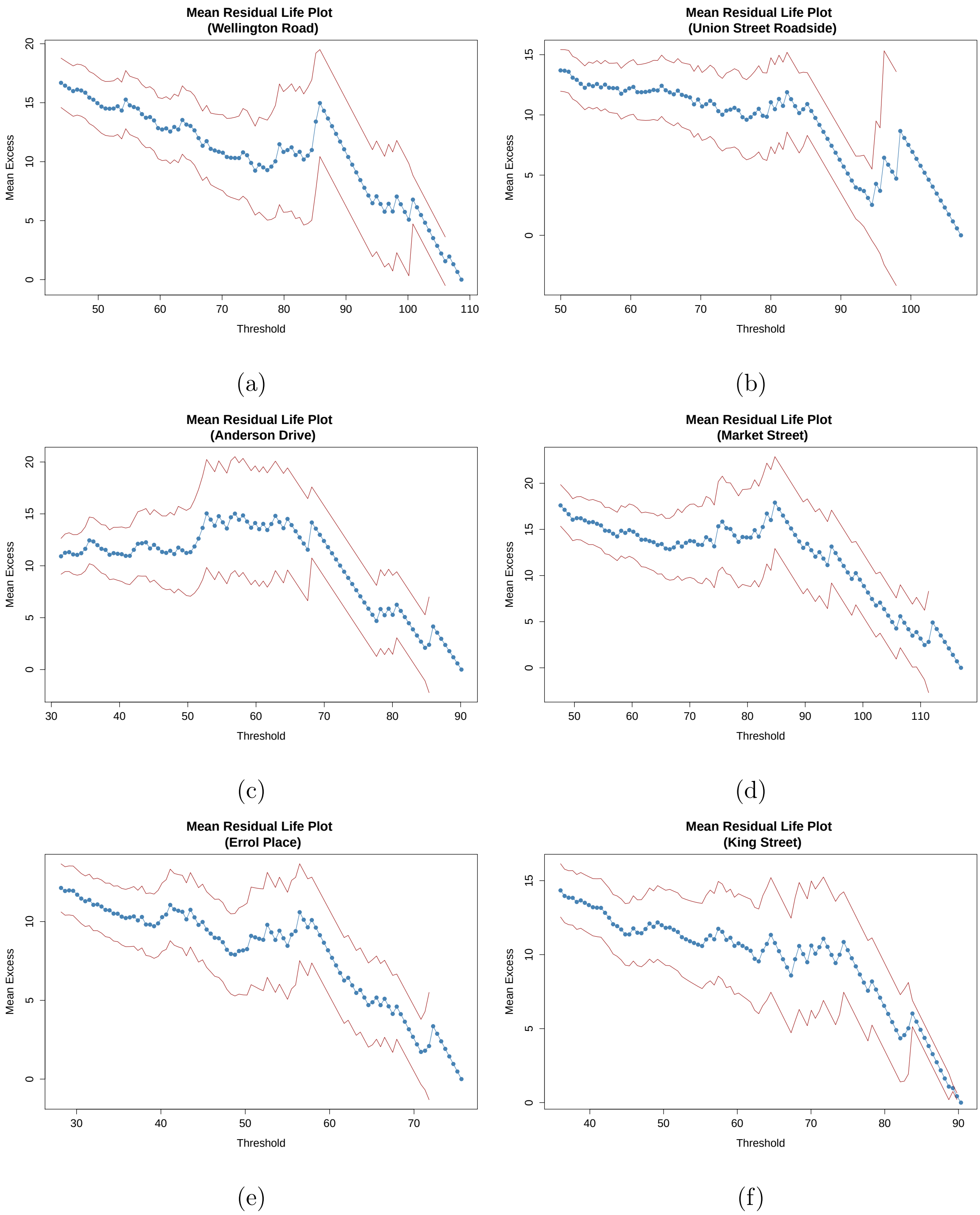


Figure 2.22: Mean Residual Life Plots for all six sites (ADMS-Urban modelled data).

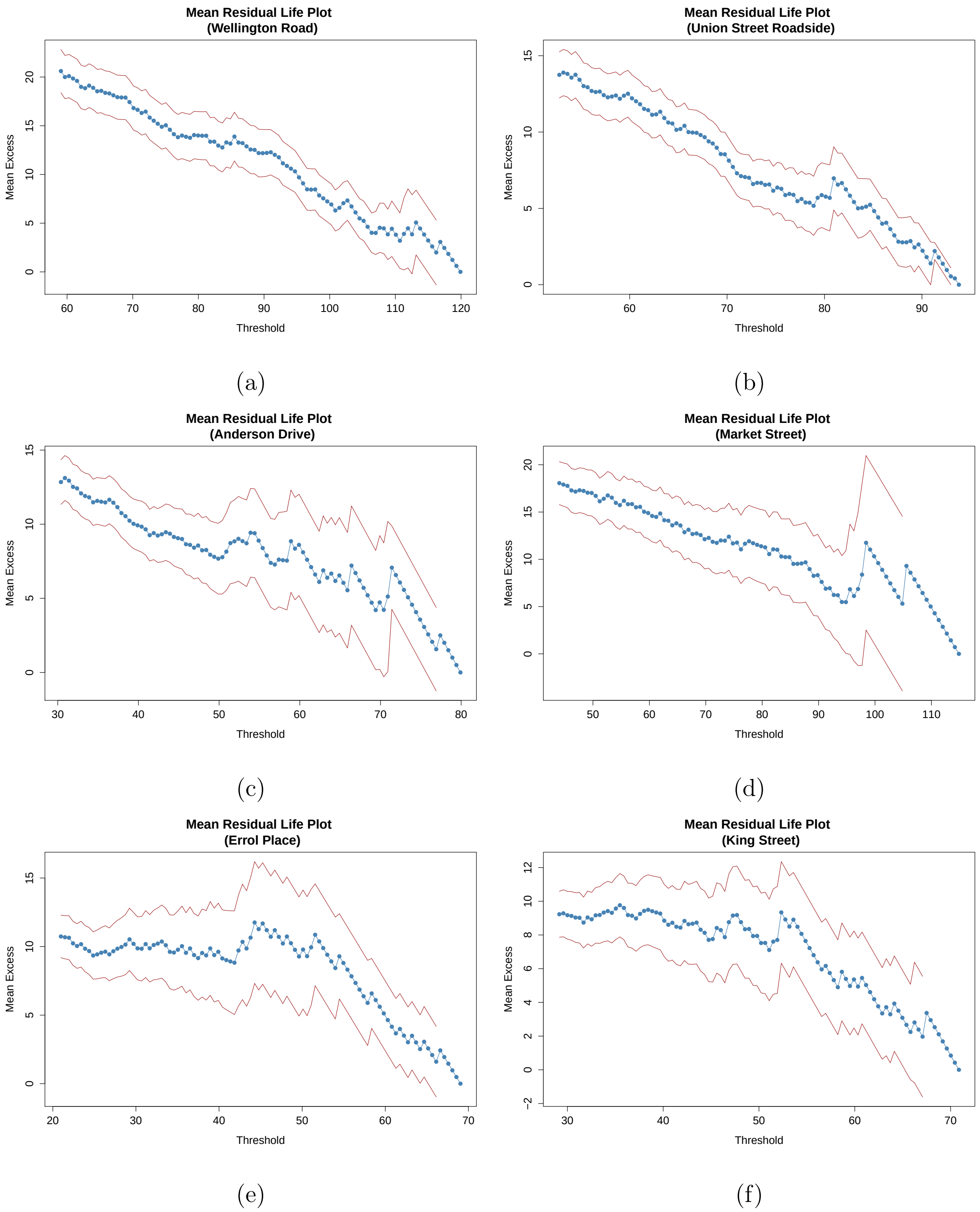


Figure 2.23: Mean Residual Life Plots for all six sites (Monitoring site data).

From Figure 2.22a, it can be seen that the mean excess is slowly gradually decreasing. This continues until around a threshold of 80 then it slightly increases again until about a threshold of 85 and then it rapidly decreases. This suggests that a threshold of 85 should be chosen as after this threshold point the mean excess is linearly decreasing. The rest of the thresholds were chosen in the same way and are presented in Table 2.10.

Table 2.10: Thresholds chosen for each of the six monitoring sites for both the modelled and monitoring data

Site	Modelled data	Monitoring data
Wellington Road	85	60
Union Street Roadside	85	80
Anderson Drive	70	60
Market Street	85	95
Errol Place	58	52
King Street	75	52

Given these thresholds have been chosen correctly, the number of events above u follows a Poisson distribution with parameter λ where λ denotes the intensity, i.e.

$$\hat{\lambda} \sim \text{Poi}(\lambda).$$

Choosing the thresholds in Table 2.10 gives the following Poisson intensity parameters in Table 2.11.

Table 2.11: $\hat{\lambda}$ values based on the thresholds chosen for each of the six monitoring sites for both the daily modelled and monitoring data

Site	Modelled data	Monitoring data
Wellington Road	7	169
Union Street Roadside	10	18
Anderson Drive	7	11
Market Street	12	6
Errol Place	10	7
King Street	7	13

From Table 2.11 it can be highlighted that for most of the monitoring sites (Wellington Road, Union Street Roadside, Anderson Drive and King Street), there appears to be more events exceeding the threshold in the monitoring NO_2 concentration data over 2012. This is due to the fact that the intensity value is larger. This highlights that the monitoring data are more variable at these sites as they are exceeding the given thresholds more and are more likely to observe and pick up on larger NO_2 concentrations than the modelled data. This suggests that the modelled data may fail to highlight NO_2 concentrations which are higher than the norm in terms of the data.

2.5 Conclusion

From this chapter it can be concluded that the ADMS-Urban modelled data and monitoring data at Wellington Road are not very well calibrated over the year 2012. This was highlighted through the daily mean NO2 concentration plot where the modelled data appeared to have lower concentrations than the monitoring data and through the difference plot where Wellington Road had the highest difference occurring with some as large as 100 μgm^{-3}. The Deming Regression suggested that for every 1 μgm^{-3} increase in the monitoring data, on average the modelled data increases by 0.4998 μgm^{-3}. In comparison to the other monitoring sites this was the poorest with the other monitoring sites having slope parameters closer to 1. It also appeared that overall, the modelled and diffusion tube data were not well calibrated at the diffusion tube locations. Deming regression emphasised that for every 1 μgm^{-3} increase in the diffusion tube data, on average the modelled data increases by 0.3863 μgm^{-3}. The bland altman plots suggested that at Wellington Road, Anderson Drive and Market Street as the variation in differences increases, the average increases. At Errol Place and King Street there appeared to be a linear relationship between the differences and the average of the modelled and monitoring data and at Union Street Roadside there appeared to be no relationship.

Despite the poor performance at Wellington Road, at the other monitoring stations the modelled and monitoring data appeared to be roughly well calibrated. At the other five sites the differences appear to be smaller than 100 μgm^{-3} and moreover, the slopes are all a lot closer to 1. This suggests that these concentrations on average at the other five sites increase at the same rate.

Furthermore, through extreme value theory and peaks over threshold it has been highlighted that taking a high threshold for example the 90^{th} percentile which was taken here that extreme concentrations given by both the monitoring and modelled data at the monitoring site Errol Place appear to follow the same pattern in terms of time. This stresses that these extreme values appear to occur at roughly the same points in time. At the other monitoring sites the exceedances of the modelled and monitoring data are not completely concurrent but the general pattern of when there is a peak occurs. It can also be noted that when the formal threshold is considered, which is found from the mean residual life plots, at the monitoring sites Wellington Road, Union Street Roadside, Anderson Drive and King Street the monitoring data has more events occurring over the chosen threshold. This highlights that the monitoring data has more variability in these cases and this can be witnessed especially at Wellington Road. The difference between the monitoring and modelled data at Wellington Road is at its highest with 7 events exceeding the threshold for the modelled data and 169 events exceeding the threshold for the monitoring data. This suggests that NO_2 concentrations which are higher than the norm in terms of the data may not be picked up on by the ADMS-Urban model.

Chapter 3

Spatial Analysis of the Monitoring Site, Diffusion Tube, DEFRA and ADMS-Urban Modelled Data

3.1 Introduction

Throughout this chapter, spatial and temporal analysis of the observed and modelled data over the year 2012 will be considered. By carrying out this analysis insight into how well the model is performing over the region of Aberdeen will be gained. To do this, monitoring site, diffusion tube and 1 km by 1 km gridded data will be aggregated to annual values. Then a statistical spatial model will be fitted and ordinary kriging will be performed to produce a spatial surface of the predicted NO_2 concentrations. The same procedure will be applied to the modelled data and these spatial surfaces will be compared. This will also be done for monthly monitoring site and diffusion tube NO_2 concentrations and again these will be compared to the spatial surface produced for the monthly modelled data. The pollutant data are logged to satisfy any normality assumptions which was not done in the analysis in the previous chapter as this analysis are partly focussed on peak values. Before going on to describe the methods, the Pollution Climate Mapping (PCM) model should be briefly described as it is the model used as the basis of the DEFRA 1 km $\times$ 1 km predictions.

As mentioned in Section 1.4.1 previously the PCM model is a group of models that are aimed in such a way to satisfy part of the UK's European Union (EU) Directive (2008/50/EC) requirements to investigate and give results on the concentrations of certain pollutants in the atmosphere (DEFRA, 2013). Basically PCM is a geographical information system (GIS) built semi-empirical model (Air Quality Modelling Review Steering Group, 2011). For more information on GIS see Esri (2015). This model is controlled by the UK National Atmospheric Emissions Inventory (NAEI) but is made up of modules which produce concentrations of various pollutants (Air Quality Modelling Review Steering Group, 2011) which were all mentioned in Chapter 1 Section 1.4.1. For more information on the UK NAEI, see the NAEI website (NAEI, 2014). In the instance of Particulate Matter (PM), different component pieces of the PM mix are produced. Computing background concentrations over the UK, where Aberdeen is the region of interest in this study, on a 1 km $\times$ 1 km grid is the foundation for the model. To obtain the regional background, measured data were used with sources that were near by modelled as area sources and large point sources. Near by sources were considered to be within around 15 km. These were modelled using a kernel approach established on ADMS 4 and large point sources were modelled directly using ADMS 4. Concentrations at roadsides are built on an empirical approach. These concentrations are described for a distance of 4 m from the kerb which is considered as an effective distance. Annual mean concentrations are provided by the PCM model, depending on empirical relationships to obtain concentrations for shorter intervals (Air Quality Modelling Review Steering Group, 2011). The maps for the years 2011 through to 2013 were downloaded for the pollutant NO_2 and the total annual mean concentrations for these 3 years were plotted over space. The plot for the year 2012 can be seen in Figure 3.1, the plot for the years 2011 and 2013 were very similar.

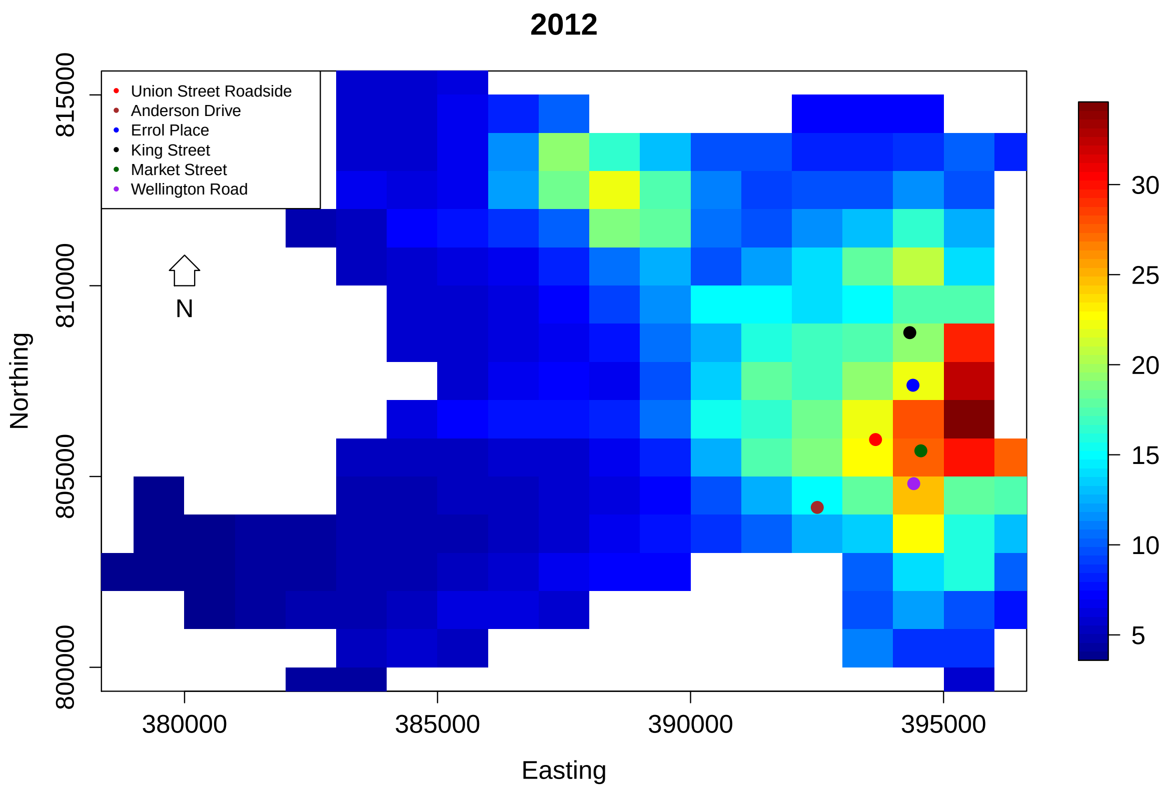


Figure 3.1: Plot of Total Annual Mean concentrations for the pollutant NO_2 over the year 2012 in Aberdeen on 1 km by 1 km grids

From Figure 3.1 it can be seen that NO_2 concentrations are higher in the east of Aberdeen where the city centre city is situated and in the northwest of the city where the airport is located. As you move further west out of the city centre the NO_2 concentrations appear lower.

Ahmadi and Sedghamiz (2007) discuss in their paper spatial and temporal analysis of groundwater level fluctuations. These data were measured monthly and analysis of 39 piezometric wells observed throughout the years of 1993 to 2004 was performed in the study area, Darab plain. To report the spatial and temporal structure of groundwater level fluctuation, geostatistics was used. Ahmadi and Sedhamiz (2007) state that in this analysis applying geostatistics, the spatial and temporal analysis were carried out on groundwater level drop and groundwater level fluctuations observed monthly throughout the study period. To explore the spatial analysis the data of water table levels of October were analysed and the difference of water table levels in the month of October 1993 and 2004 was utilized as evidence of the calculated groundwater level decrease (12.6 m on average). It was concluded by Ahmadi and Sedhamiz (2007) that geostatistics can report stochastic structure of groundwater level differences both in space and time. Ordinary kriging with cross validation was carried out and this produced adequate estimations of groundwater level drop in points that are apparently unspecified. Through carrying out temporal analysis it was emphasised that groundwater fluctuations also have temporal structure. This time universal kriging with cross validation was carried out and this produced extremely adequate estimations of groundwater level in a succession of observed groundwater levels (Ahmadi and Sedghamiz, 2007).

Wong *et. al* (2004) investigate respiratory impacts in children by evaluating the role of exposure to surrounding air pollutants as risk factors. Wong *et. al* (2004) particularly explain an effort using Environmental Protection Agency's Aerometric Information Retrieval System monitoring data to evaluate O_3 and PM_{10} concentrations. These pollutant concentrations are evaluated at census block groups and these have been restricted to countries that have been seen by National Health and Nutrition Examination Survey-III and four individual interpolation techniques have been implemented to the monitoring data to obtain air concentration levels. Wong *et. al* (2004) then investigate approach-specific differences in concentration levels and they establish conditions under which each individual approach yield notably different concentration values. They conclude that in the majority of the US where monitor density was considerably small, different interpolation approaches do not yield substantially different approximations. Moreover, in parts of the US where monitor density was considerably big, Wong *et. al* (2004) discovered significant differences in exposure approximations over the interpolation approaches (Wong *et. al*, 2004).

Matejicek (2014) carry out various geostatistical methods to predict the pollutants NO_2 and PM_{10}. Their main data for geostatistical approaches derive from sample points that are produced from a system of automatic monitoring stations. Furthermore, other sample points evaluated by geographically weighted regression accompany these data. To examine spatially differing associations between air pollution, as a response variable and various independent variables, geographically weighted regression is carried out to produce a local form of linear regression. Matejicek (2014) uses methods for spatial interpolation that are established on geostatistical approaches such as ordinary kriging. Matejicek (2014) carries out the analysis on Prague (capital of the Czech Republic) and geographically weighted regression and the prediction maps of NO_2 and PM_{10} highlight extremely unprotected sites that strongly suggest the need for urgent measures in urban air quality and traffic management (Matejicek, 2014).

3.2 Methodology

The following method described is based on information given by Diggle and Ribeiro Jr. (2007). A geostatistical process is used here as the spatial area which will be denoted by A is a continuous 2-dimensional region, $A \subset \mathbb{R}^2$. A set number of locations are observed in practice and these are selected by the person collecting the data.

Prior to fitting a spatial statistical model and performing ordinary kriging to produce a spatial surface of the data, there will be a brief introduction to the method involved. Diggle and Ribeiro Jr. (2007) state that a fundamental geostatistical model includes two or more components. The first component is a real-valued stochastic process which is defined as:

$$\{S(\mathbf{x}) : \mathbf{x} \in A\},$$

where A is some spatial region. This is usually supposed to be a partial realisation of the stochastic process given by:

$$\{S(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2\},$$

over the full plane. The second component is a multivariate distribution for the random variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$ conditional on $S(.)$. $S(\mathbf{x})$ denotes the signal and Y_i denotes the response. A is a subset of 2-dimensional space which can't be changed. Data can take place at any location $\mathbf{x} = (x_1, x_2)$ in the spatial region A, although practically data are measured at a fixed number of locations (Diggle and Ribeiro Jr., 2007). In this study the coordinates $\mathbf{x} = (x_1, x_2)$ are easting and northing in metres.

3.2.1 Parameter Estimation

Diggle and Ribeiro Jr. (2007) consider methods for producing an adequate geostatistical model and evaluating its parameters. The linear Gaussian model will be looked at, parameter estimation through using maximum likelihood estimation could also be explored but in this study it was found that the linear Gaussian model was adequate. For more information on maximum likelihood estimation see Diggle and Ribeiro Jr. (2007). Consider for data $Y_i : i = 1, \ldots, n$ measured at spatial locations $x_i : i = 1, \ldots, n$ where the data has a mean structure $\mathbb{E}[Y_i] = \mu_i$ and covariance structure that has to be decided. It is supposed that $\mu_i = \mu(x_i)$ where

$$\mu(x) : \beta_0 + \sum_{j=1}^{p} \beta_j d_j(x).$$

From the above equation the $\mathrm{d}_j(x)$ denotes the spatial explanatory variables, β_0 denotes the intercept and β_j denotes the mean parameters to be estimated. To begin with the ordinary least squares criterion is used to estimate the mean parameters. This is done by picking estimates to minimise the residual sum of squares defined as:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \mu_i)^2.$$

By minimising this quantity the estimates $\hat{\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\beta}} = (D^T D)^{-1} D^T Y,$$

where $Y = (Y_1, \ldots, Y_n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and D is a matrix made up of firstly a column of ones and the rest of the columns are made up of the values of the explanatory variables. This matrix has dimensions n by $(p + 1)$. Once the $\hat{\boldsymbol{\beta}}$ have been achieved, the residuals $R_i : 1, \ldots, n$ are given by components of the vector

$$R = Y - D\hat{\boldsymbol{\beta}}.$$

These residuals are used to find an adequate parametric model for the covariance structure and to get starting off estimates of covariance parameters (Diggle and Ribeiro Jr., 2007). Diggle and Ribeiro Jr. (2007) then proceed on to discuss variograms and this will be explored and explained in the following subsection.

3.2.2 Variograms and Correlation Functions

Diggle and Ribeiro Jr. (2007) state that the theoretical variogram of a spatial stochastic process is given by

$$V(\mathbf{x}, \mathbf{x}') = \frac{1}{2} \mathrm{Var}\{S(\mathbf{x}) - S(\mathbf{x}')\}$$

If the process is stationary or intrinsic, the variogram decreases to

$$\mathbf{u} = ||\mathbf{x} - \mathbf{x}'||.$$

This means that the covariance function or the variogram describe the second-moment properties of a stationary stochastic process S($\mathbf{x}$). In the case of a stationary process where the mean is uniform the variogram can also be given by

$$V(\mathbf{u}) = \frac{1}{2}\mathbb{E}[\{S(\mathbf{x}) - S(\mathbf{x}-\mathbf{u})\}^2].$$

The usual variogram is an increasing function which is monotonic with the upcoming characteristics which are listed below:

- The intercept denoted by τ^2 represents the nugget variance.
- The asymptote which will be denoted by $\tau^2 + \sigma^2$ represents the sill (Var(Y)) where σ^2 denotes the signal variance.
- The correlation function $\rho(\mathbf{u})$ controls the process taken for the variogram to increase from τ^2 to $\tau^2 + \sigma^2$.
- The range of the variogram is defined as the value when $\rho(\mathbf{u}) = 0$ for $\mathbf{u}$ bigger than any finite number. The range is undefined if the correlation function only approaches 0 asymptotically as $\mathbf{u}$ gets bigger.

The graphic representation of a classic variogram is indicated in Figure 3.2 below.

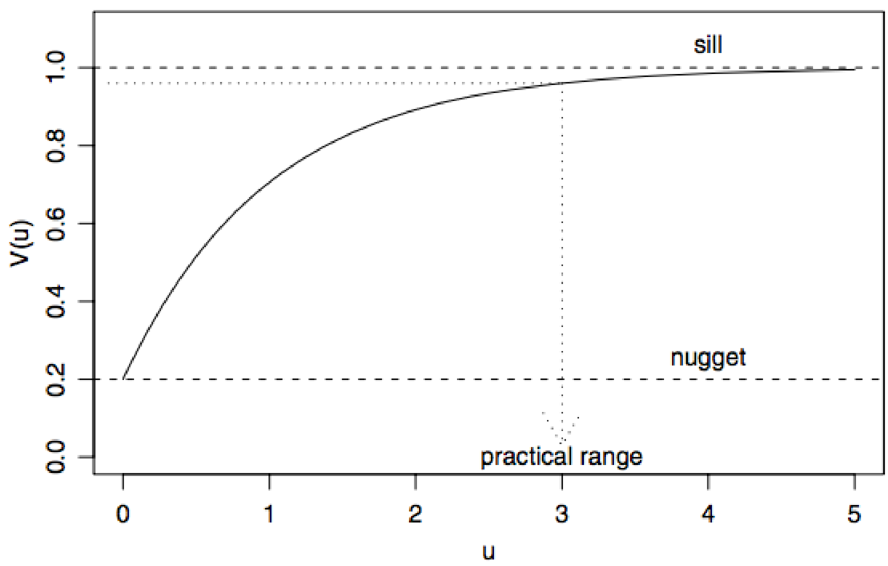


Figure 3.2: Graphic representation of a classic variogram, with structural parameters specified (Diggle and Ribeiro Jr., 2007)

Figure 3.2 highlights the characteristics of the variogram and that the variogram is an monotonic increasing function.

The most general correlation function, The Matérn Family, will now be introduced and this was the correlation function used in this study. This family of correlation functions works very nicely as it covers the following two demands:

- The correlation between $S(\mathbf{x})$ and $S(\mathbf{x}')$ gets smaller as the distance $\mathbf{u} = ||\mathbf{x}$ - $\mathbf{x}'||$ gets bigger and,
- When considering the fundamental spatial process $S(\mathbf{x})$ differing applications may present degrees of smoothness which are not the same.

The correlation function is given by,

$$\rho(\mathbf{u}) = \{2^{k-1}\Gamma(k)\}^{-1}(\mathbf{u}/\phi)K_k(\mathbf{u}/\phi),$$

where K_k represents a modified Bessel function of order k, ϕ represents a scale parameter which is greater than 0 and has the dimensions of $\mathbf{u}$ and the order k is also greater than 0 and represents the shape parameter. The analytic smoothness of the fundamental spatial process $S(\mathbf{x})$ is controlled by the shape parameter ϕ (Diggle and Ribeiro Jr., 2007).

3.2.3 Exploring the presence of spatial correlation in data

One way to examine if there is spatial correlation in a data set is to evaluate the empirical semi-variogram. Let

$$N(u) = \{(\mathbf{x}_i, \mathbf{x}_j) : ||\mathbf{x}_i - \mathbf{x}_j|| = u\}$$

represent the set of pairs of spatial locations at a distance u apart. Let the size or amount of points in this set be represented by $|N(u)|$. The empirical semi-variogram at distance u is given by

$$\hat{\gamma}_Y(u) = \frac{1}{2|N(u)|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in N(u)} [y(\mathbf{x}_i) - y(\mathbf{x}_j)]^2.$$

The amount of points in $N(h)$ may be one for many visible distances u since the original data points are not evenly positioned all over the spatial region A. Thus, there may not be a sufficient amount of points to take the mean of to produce a good approximation of the true variogram. Therefore the binned empirical semi-variogram is considered.

Assume the space of distances are divided into K intervals where the intervals represent the bins

$$I_k = (u_{k-1}, u_k],\ k = 1, \ldots, K \text{ where } 0 = h_0 < h_1 < \cdots < h_K.$$

Let the midpoint of the interval be represented by

$$u_k^n = \frac{(u_{k-1} + u_k)}{2}.$$

The pairs of distances in all intervals are computed and

$$\hat{\gamma}_Y(u_k^n) = \frac{1}{|2N(u_k)|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in N(u_k)} [y(\mathbf{x}_i) - y(\mathbf{x}_j)]^2$$

denotes the binned empirical semi-variogram (Diggle and Ribeiro Jr., 2007).

Monte-carlo envelopes

One way to determine whether there is spatial correlation is to plot the semi-variogram. On this plot the upper limits and lower limits for the set of semi-variograms that would have occurred under independence is added. Both of these limits are calculated using Monte Carlo approaches and are frequently referred to as Monte Carlo envelop. To assess evidence of spatial correlation, we observe that the evaluated semi-variogram from the data is entirely contained within the envelop (Diggle and Ribeiro Jr., 2007).

3.2.4 Spatial Prediction

A primary aim of a geostatistical analysis is to estimate the process at locations $\mathbf{x}_0$ which have not been measured. This can be achieved for a regular grid of locations that have not been measured and by doing so a map can be created of the fundamental process for the region being studied. The main issue here is that based on the data $\mathbf{s}$ $= (s(\mathbf{x}_1), \ldots, s(\mathbf{x}_n))^T$, it is of interest to find the optimum prediction of $S(\mathbf{x}_0)$ at a new location $\mathbf{x}_0$.

This is not an easy issue but it can be made easier by requiring the prediction of $S(\mathbf{x}_0)$ to be linear in the observed data. This is given by

$$P_{\mathbf{s}}(\mathbf{x}_0) = a_0 + \sum_{k=1}^{n} a_k S(\mathbf{x}_k). \tag{3.1}$$

From Equation 3.1 a_0 denotes some constant, the prediction weights are denoted by $\mathbf{a}$ $= (a_1, \ldots, a_n)$ and the linear predictor operator is given by P (Diggle and Ribeiro Jr., 2007).

Kriging

Kriging is a widely used approach for prediction. This approach is based on obtaining the Best Linear Unbiased Prediction (BLUP) for $S(\mathbf{x}_0)$ given data $\mathbf{s} = (s(\mathbf{x}_1), \ldots, s(\mathbf{x}_n))^T$. To achieve the BLUP, $(a_0, \mathbf{a})$ are picked such that they minimise the mean squared prediction error. This quantity is given by

$$\text{MSPE} = \mathbb{E}[(S(\mathbf{x}_0) - P_{\mathbf{s}}(\mathbf{x}_0))^2],$$

where the prediction error at $\mathbf{x}_0$ can be represented by

$$U(\mathbf{x}_0) = S(\mathbf{x}_0) - P_{\mathbf{s}}(\mathbf{x}_0).$$

The minimiser of this can be found to be $\mathbb{E}[S(\mathbf{x}_0)|\mathbf{S}]$. This is the conditional expectation of the process at the locations $\mathbf{x}_0$ which have not been measured given the observed data.

Lets assume that the data $\mathbf{s}$ has a mean which is not constant i.e. $\mathbb{E}(\mathbf{S}) = D\boldsymbol{\beta}$ so the data $\mathbf{s} \sim \text{N}(D\boldsymbol{\beta}, \Sigma(\boldsymbol{\theta}))$ where $\Sigma(\boldsymbol{\theta})$ is an n by n covariance matrix for the n observations and $\boldsymbol{\theta} = (\sigma^2, \tau^2, \phi)$ which express the partial sill, nugget and range of the correlation structure respectively. Then merging the data $\mathbf{S}^* = (S(\mathbf{x}_0), \mathbf{S})$ gives the distribution

$$\mathbf{S}^* = \begin{pmatrix} S(\mathbf{x}_0) \\ \mathbf{S} \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mathbf{d}_0\boldsymbol{\beta} \\ D\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} C_S(\mathbf{0}, \boldsymbol{\theta}) & \mathbf{c}_S(\mathbf{x}_0, \boldsymbol{\theta})^T \\ \mathbf{c}_S(\mathbf{x}_0, \boldsymbol{\theta}) & \Sigma(\boldsymbol{\theta}) \end{pmatrix}\right)$$

where the vector of covariates is denoted by $\mathbf{d}_0$ at the location that has not been observed $\mathbf{x}_0$ and $C_S(.)$ represents a covariance function. Then

$$S(\mathbf{x}_0)|\mathbf{S} \sim \text{N}(\mathbb{E}[\widehat{S(\mathbf{x}_0)}|\mathbf{S}], \text{Var}[\widehat{S(\mathbf{x}_0)}|\mathbf{S}]). \tag{3.2}$$

From Equation 3.2,

$$\mathbb{E}[\widehat{S(\mathbf{x}_0)}|\mathbf{S}] = \mathbf{d}_0\hat{\boldsymbol{\beta}} + \mathbf{c}_S(\mathbf{x}_0, \hat{\boldsymbol{\theta}})^T \Sigma(\hat{\boldsymbol{\theta}})^{-1}(\mathbf{S} - D\hat{\boldsymbol{\beta}})$$
$$\text{Var}[\widehat{S(\mathbf{x}_0)}|\mathbf{S}] = C_S(\mathbf{0}, \hat{\boldsymbol{\theta}}) - \mathbf{c}_S(\mathbf{x}_0, \hat{\boldsymbol{\theta}})^T \Sigma(\hat{\boldsymbol{\theta}})^{-1} \mathbf{c}_S(\mathbf{x}_0, \hat{\boldsymbol{\theta}}).$$

This is known as the Universal Kriging predictor (Diggle and Ribeiro Jr., 2007).

3.3 Results

Firstly, observed annual data will be investigated in great detail and then kriging results of modelled annual data and monthly observed and modelled data will be produced.

The year 2012 will be focused on as mentioned previously. In this study, once the DEFRA data, monitoring site and diffusion tube data were aggregated annually over the year 2012 and transformed into the form of geodata in R, Figure 3.3 was produced. For this analysis the DEFRA modelled results have became points by attributing the concentration to the centroid of the grid cell. Figure 3.3 was used initially to give some impression of what the annual observed data looked like over the region of Aberdeen in 2012. For this analysis, the following packages in *R geoR* (CRAN, 2015a), *gstat* (CRAN, 2016a) and *sp* (CRAN, 2016b) were used.

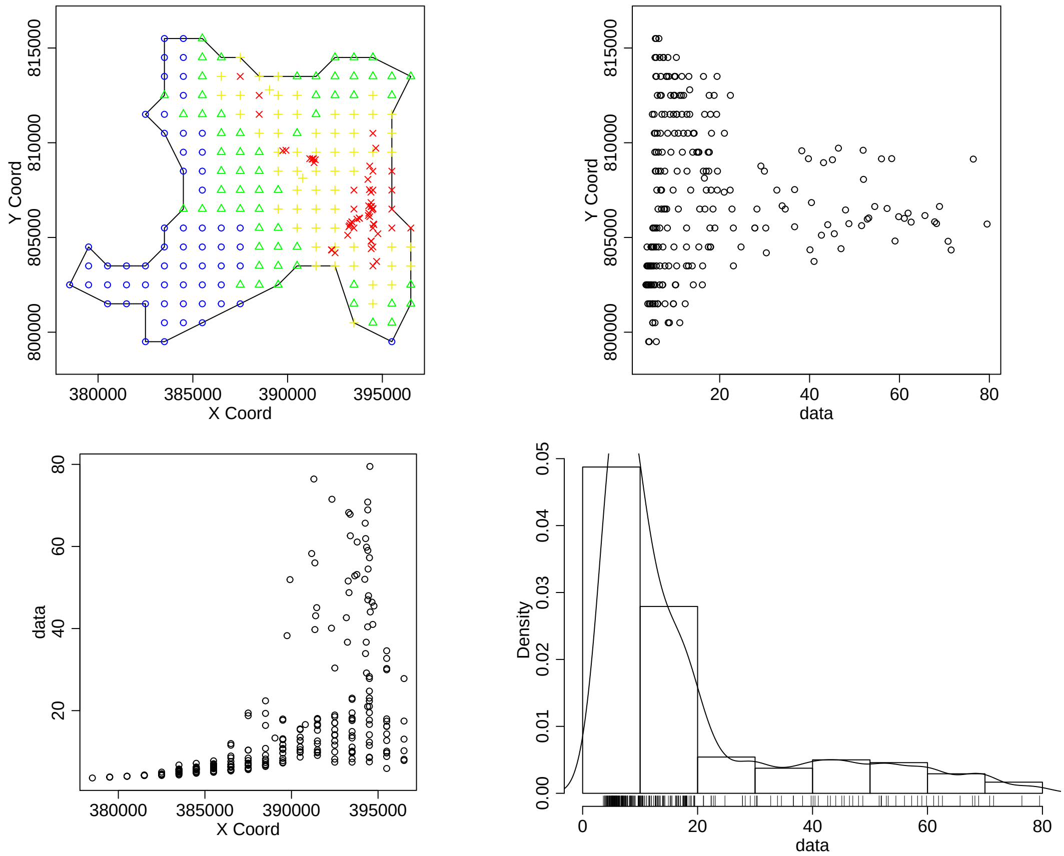


Figure 3.3: Plots to give an initial impression of what the data looks like

The top left plot in Figure 3.3 highlights that in the east and northwest of Aberdeen the NO_2 concentrations are at their highest as highlighted by the red crosses. This is hardly surprising as this is where the city centre and the airport are situated in Aberdeen. As you move further west of the city the concentrations appear to get lower and this is indicated by the green triangles and blue circles. The top right plot in Figure 3.3 represents that over space data observations greater than 20 μgm^{-3} are contained in one small area between roughly 805000 m and 810000 m. The bottom left plot in Figure 3.3 shows that again over space data observations greater than 20 μgm^{-3} are contained in one small area between roughly 390000 m and 395000 m. Finally the bottom right plot in Figure 3.3 highlights that the data is skewed to the right and a log transformation may be needed. Logging the NO_2 data produces the following Quantile-Quantile (Q-Q) plot,

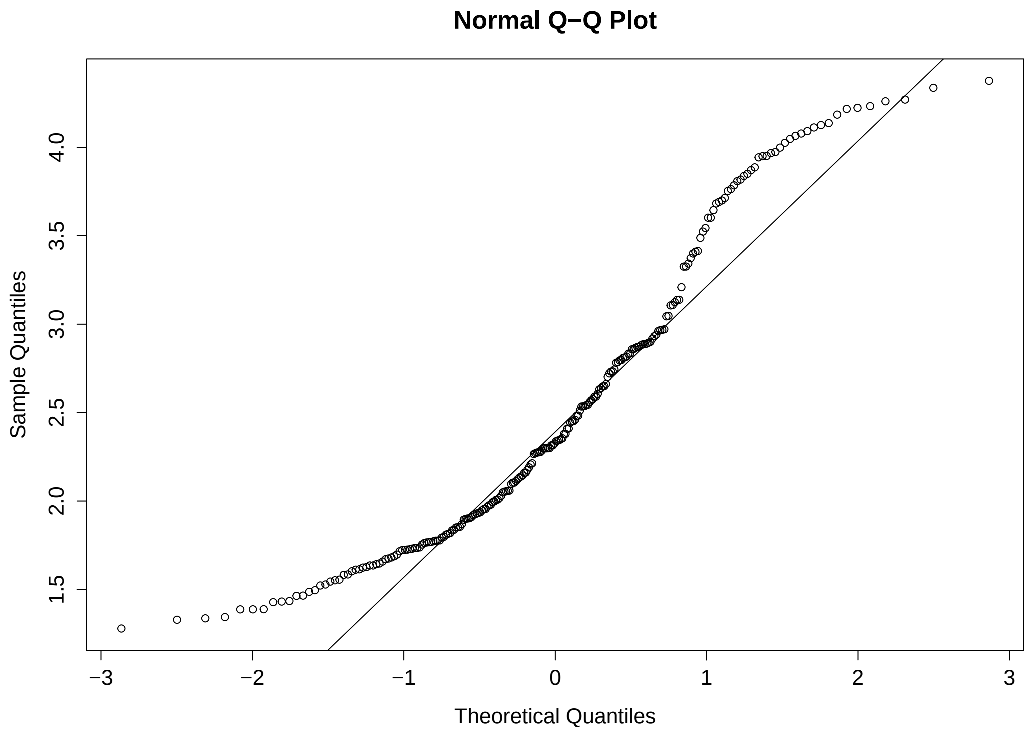


Figure 3.4: Q-Q Plot of log(NO_2) concentrations

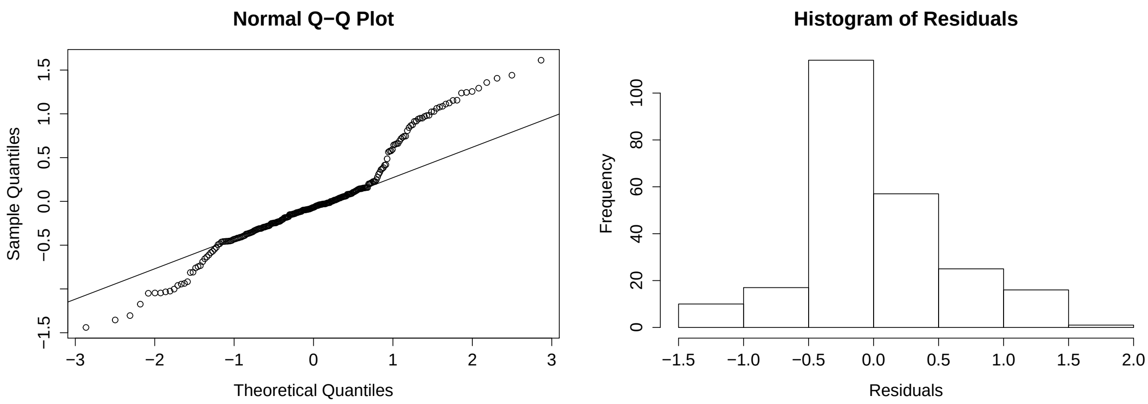
From Figure 3.4 it can be seen that logging the NO_2 concentrations doesn't give a perfect normal density and that the tails are very heavy. In terms of the observed data, this could be because monitoring site, diffusion tube and DEFRA data have been combined together. After carrying out a linear model where the log(NO_2) concentrations from the DEFRA, diffusion tube and monitoring site data denote the response variable and Easting and Northing denote the explanatory variables the following results were produced,

Table 3.1: Results from linear model

Coefficient	Estimate	Standard Error	P-Value
Intercept	-55.34	7.859	<0.0001
Easting	0.0001334	0.000008156	<0.0001
Northing	0.000007261	0.000009071	0.424

From Table 3.1 it can be observed that from this linear model a clear easting effect is present but no northing effect. However, the northing effect should be kept in the model to provide a rotationally invariant surface. Fitting this linear model gave a R^2 (adjusted) value of 52.84%. This highlights that 52.84% of the variability in the data is explained by this linear model, while adjusting for all the parameters in the model. This value of 52.84% suggests there is still room for improvement in spatial predictions, although considering the complexity of the problem, the value is not discouraging. It can also be stated from this model that for every 1 unit towards the Easting direction, on average log(NO_2) concentration increases by 0.0001334 units and for every 1 unit towards the northing direction, on average log(NO_2) concentration increases by 0.000007261 units.

Checking Assumptions for the Linear model fitted



(a) Normal Quantile-Quantile (Q-Q) Plot of the residuals

(b) Histogram of the residuals

Figure 3.5: Residual plots

From Figure 3.5a, it can be seen that the points seem to drift away from the straight line at the beginning and very much so towards the end suggesting that the upper tail may have to be investigated. Figure 3.5b highlights that the histogram of residuals is looking to be fairly normal, however, the tails again especially the upper tail may have to be further examined.

Variogram

Figure 3.6 represents Monte Carlo envelopes for the variogram of the residuals from the simple linear model of the log(NO_2) concentration data.

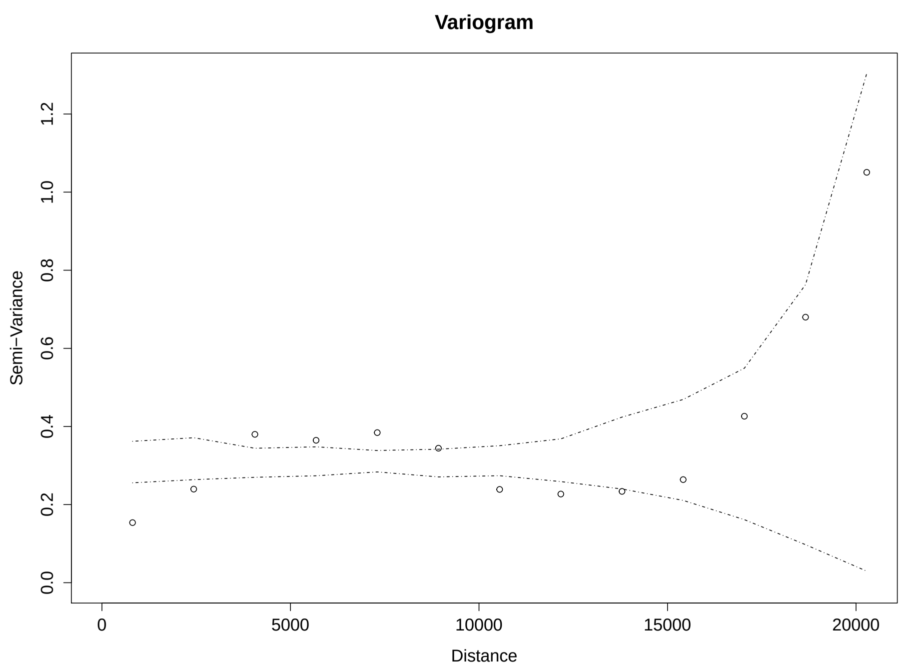
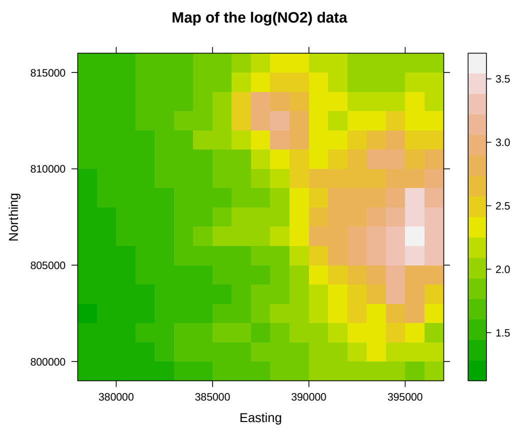


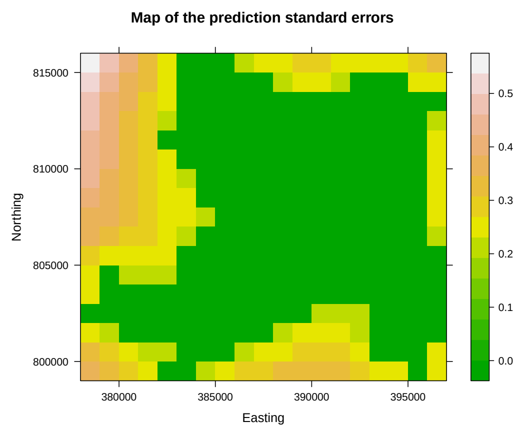
Figure 3.6: Monte Carlo envelopes for the variogram of residuals after fitting the simple linear model

This seems to highlight that there may be spatial correlation present as some of the points are outside the monte carlo envelopes. However, further investigation highlighted that the linear model is adequate for the data in this case. Then the next step was to use ordinary kriging to predict the log(NO_2) concentrations across the region of Aberdeen with a Matérn model.

Results from using Ordinary Kriging



(a)

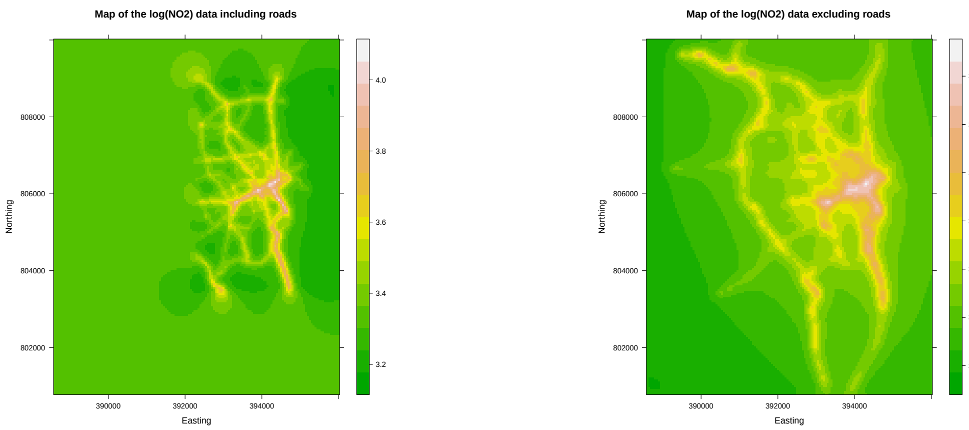


(b)

Figure 3.7: Map of the annual kriged observed log(NO_2) data and map of the observed prediction standard errors for 2012

From Figure 3.7a it shows that NO_2 concentrations are higher in the east of Aberdeen and up towards the northwest. Once more this is expected as the city centre is in the east of the city and the airport is up near the northwest of the city. It is also predicted that as you move further away from the city centre, further west, that the NO2 concentrations get lower. Figure 3.7b highlights that the predicted standard errors appear to be uniformly small over the region of Aberdeen for the year 2012. The green region in Figure 3.7b is where the data points are to be predicted and outwith this region there is no data to be predicted.

This same procedure was carried out for the annual ADMS-Urban modelled data and the monthly observed and ADMS-Urban modelled data. When carrying out this procedure for the modelled data the main city centre of Aberdeen was focussed on and 7454 pixels in the modelled data were examined. These are the pixels that occupy the main city centre of Aberdeen and were the monitoring site and diffusion tubes are located. The main reason and benefit of doing this was to resolve computational challenges with the scale of the data and model fitting. The modelled data was first of all investigated with the gridded pixels representing the roads included and then these pixels were removed to see the impact the roads had on the kriging predictions. Figure 3.8 highlights the result when ordinary kriging is used to predict the $\log(NO_2)$ concentrations for the annual modelled data including and excluding the roads.



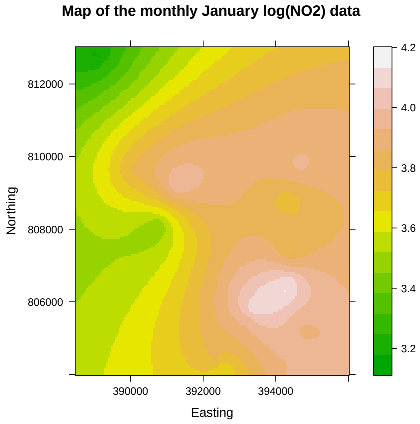
(a) Including roads

(b) Excluding roads

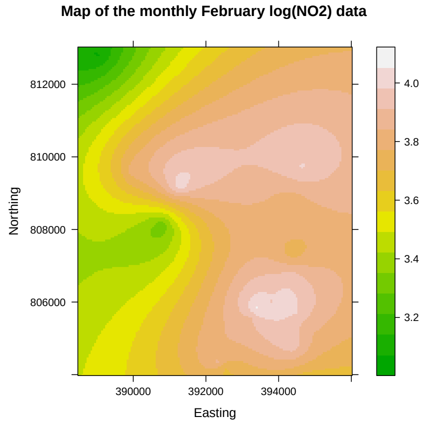
Figure 3.8: Map of the annual kriged modelled $\log(NO_2)$ data

In both Figures 3.7a and 3.8 there appears to be high concentrations occurring in the east of Aberdeen where the city centre is located as mentioned. This highlights that the model shows that the city centre of Aberdeen is highly polluted with NO_2 concentrations. From Figures 3.8a and 3.8b the importance of the roads in Aberdeen can be observed as even when these pixels have been removed, the spatial structure of the roads is still present. It is highlighted in Appendix B that the standard errors for the annual modelled data appear to be small in both cases over the year 2012.

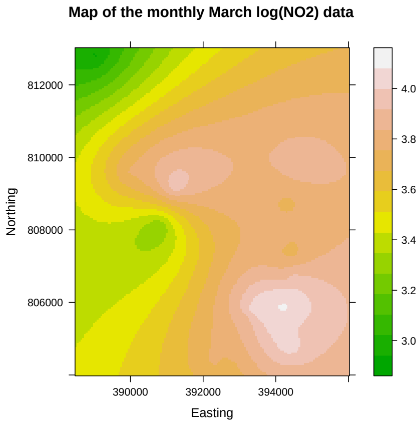
The diffusion tube and monitoring site aggregated monthly data will now be examined. The same procedure that took place for the annual data was carried out for the monthly data. Using ordinary kriging to predict the $\log(NO_2)$ concentrations for each month yields the following results,



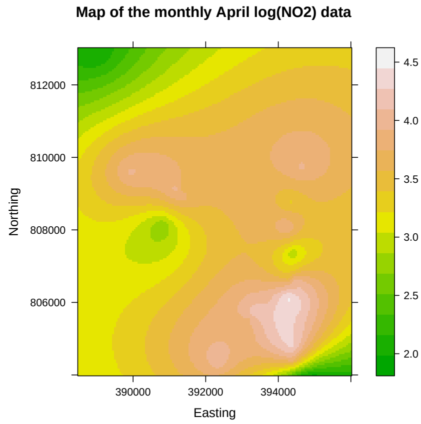
(a)



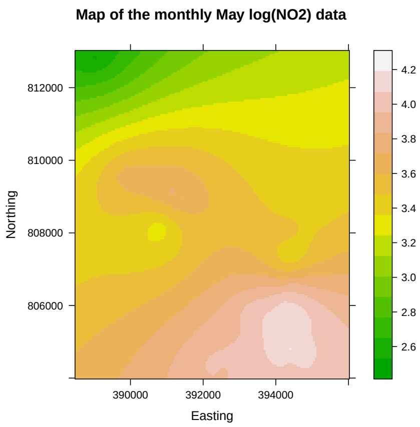
(b)



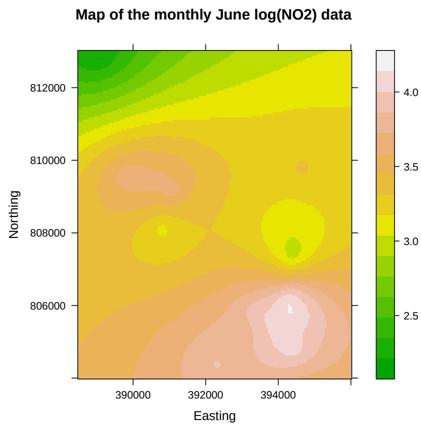
(c)



(d)



(e)



(f)

Figure 3.9: Map of the monthly kriged observed NO_2 data (January to June)

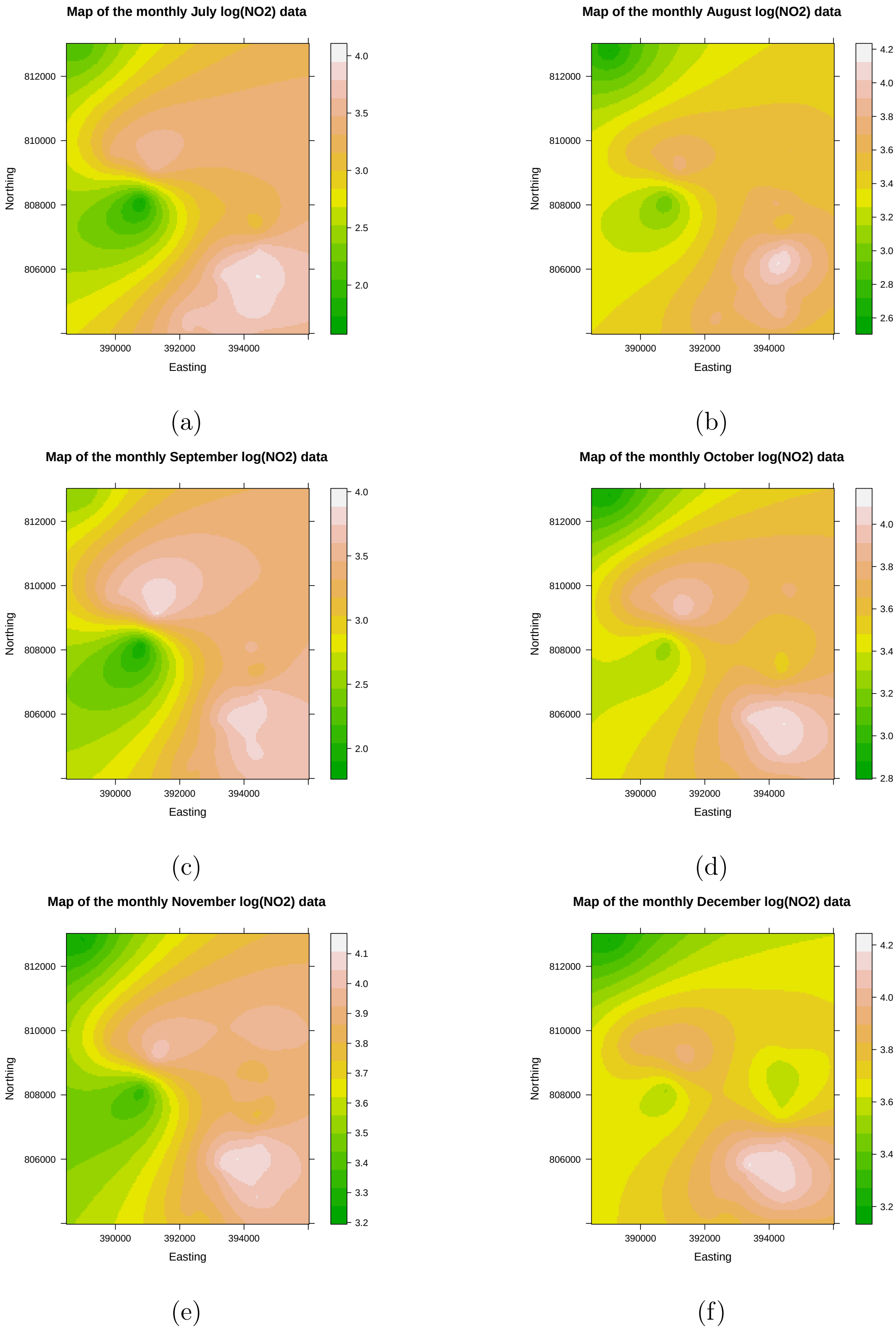


Figure 3.10: Map of the monthly kriged observed NO_2 data (July to December)

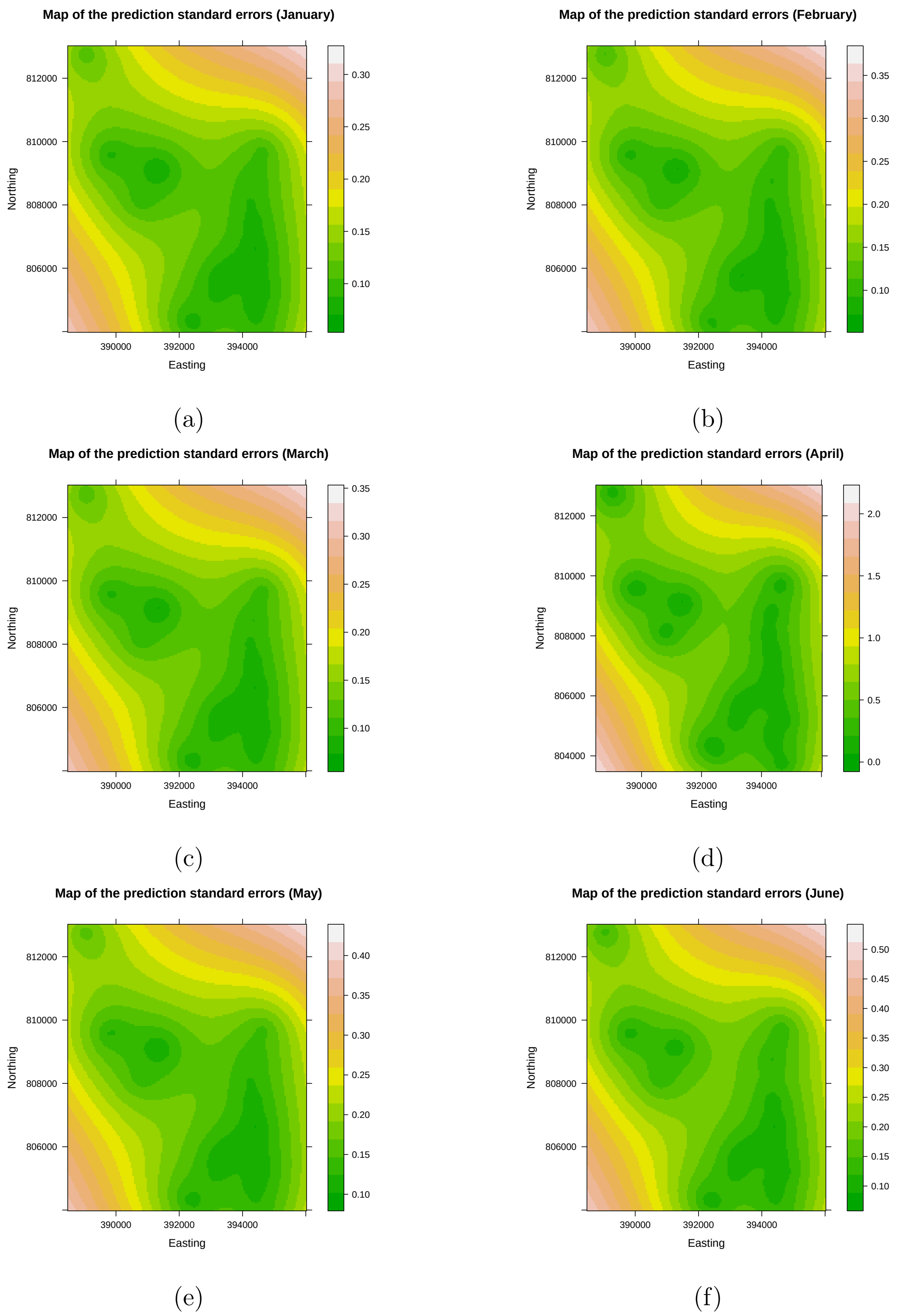


Figure 3.11: Map of the observed prediction standard errors (January to June)

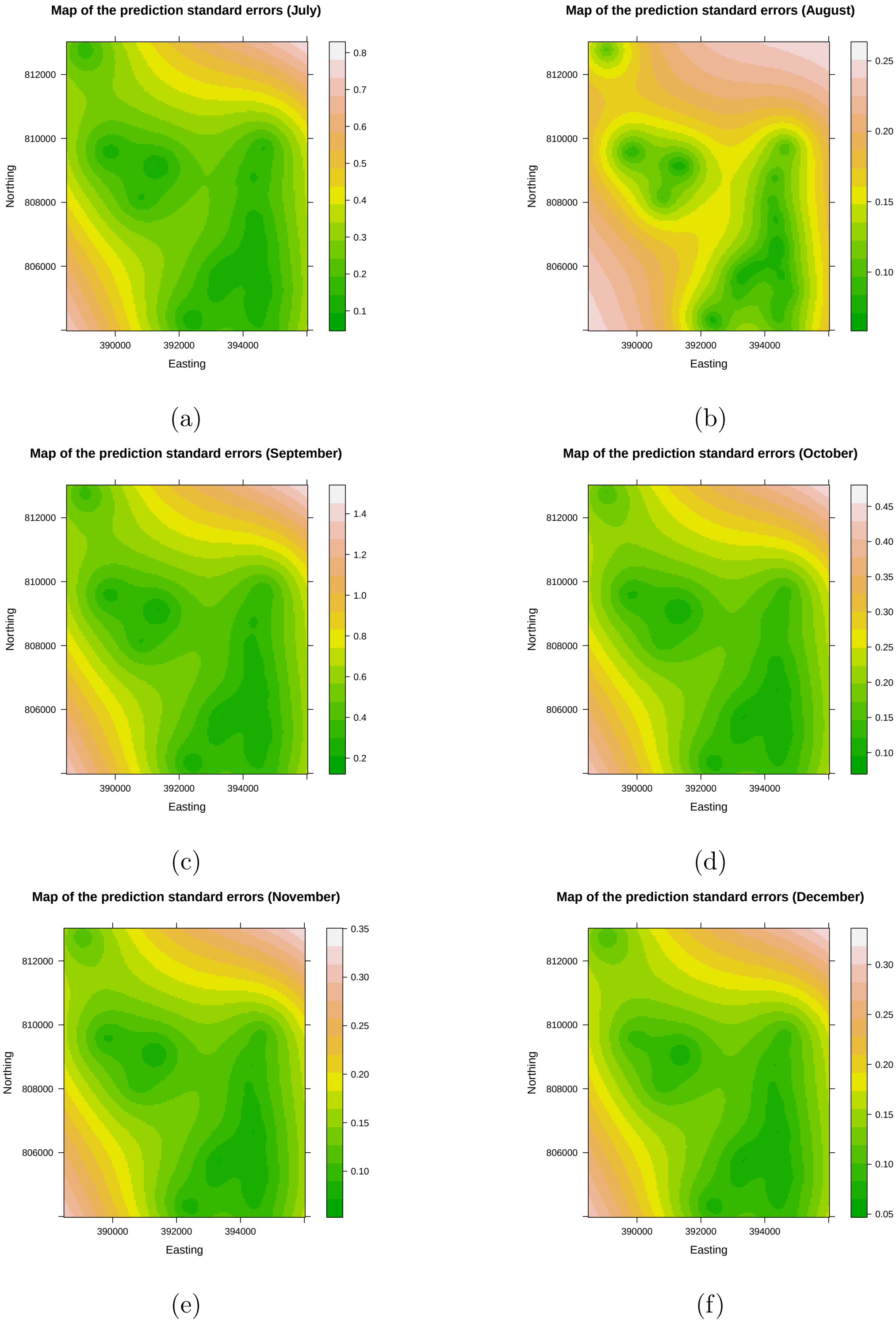
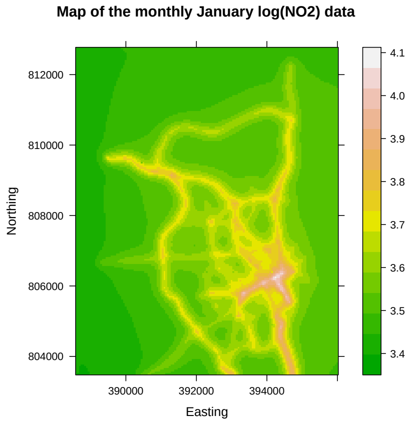
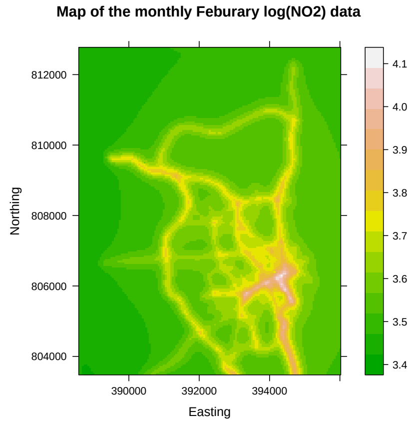


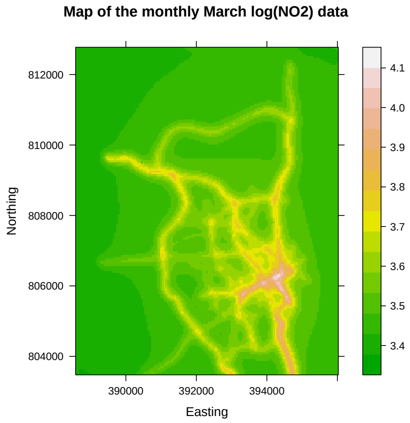
Figure 3.12: Map of the observed prediction standard errors (July to December)



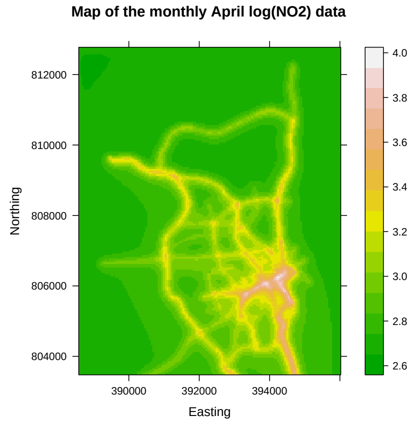
(a)



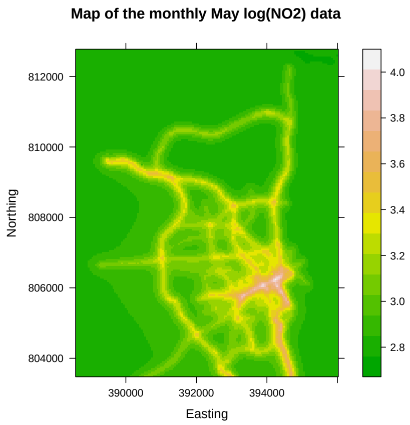
(b)



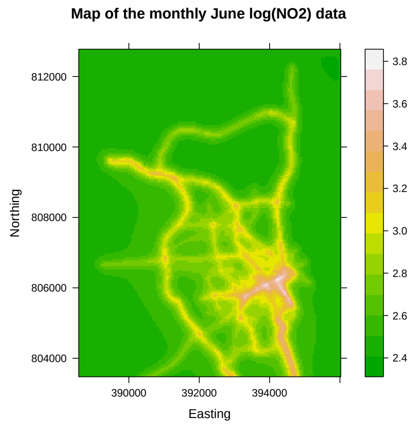
(c)



(d)



(e)



(f)

Figure 3.13: Map of the monthly kriged modelled NO_2 data (January to June)

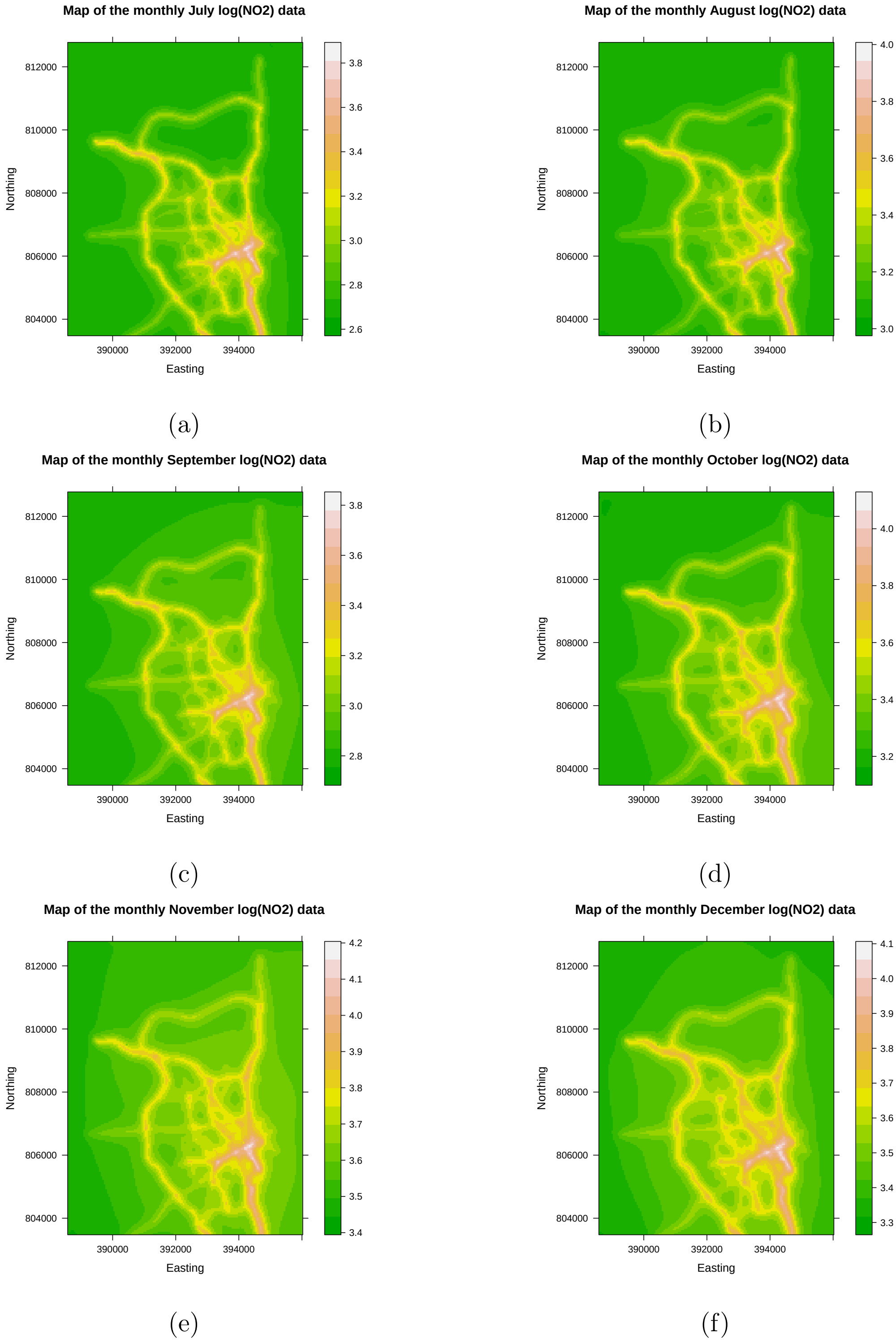
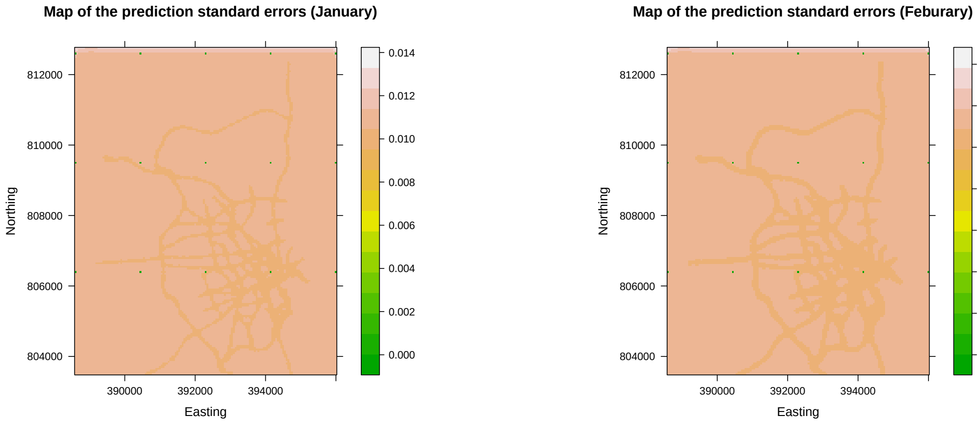


Figure 3.14: Map of the monthly kriged modelled NO_2 data (July to December)



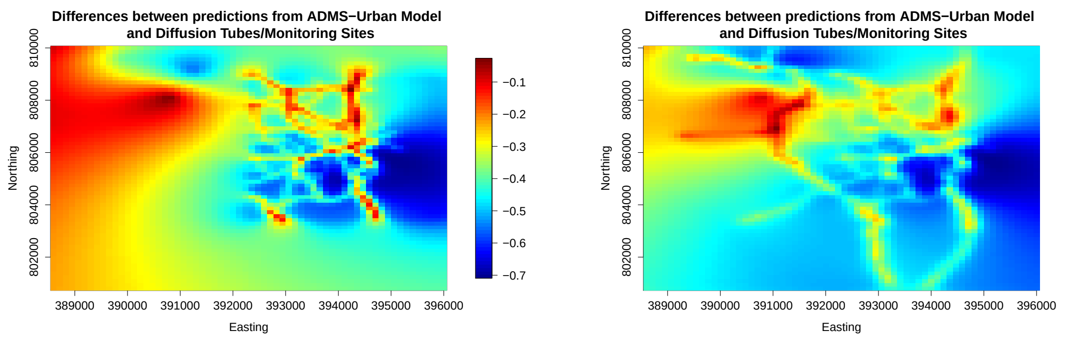
(a) (b)

Figure 3.15: Map of the modelled prediction standard errors (January and February)

Before discussing the following figures, it should be noted that Figures 3.9, 3.10, 3.13 and 3.14 do not represent the whole of Aberdeen but have been zoomed into the city centre and airport region of Aberdeen where the diffusion tubes and monitoring sites are situated. From Figures 3.9, 3.10, 3.13 and 3.14 it can be suggested that there doesn't appear to be much change in the NO_2 concentrations over the months of 2012. From these Figures it can also be observed that there always appears to be high concentrations occurring in the southeast of the area being zoomed into and this unsurprisingly happens to be around the city centre region reinforcing the NO_2 concentrations are higher there. This highlights that the model is successfully capturing where the NO_2 concentrations are at their highest. It can also be observed from Figures 3.13 and 3.14 that there is details in the modelled data that we don't see in the observed data since the modelled data is of such a high spatial resolution. Figures 3.11 and 3.12 emphasises that for most of the months the predicted standard errors appear to be small over the diffusion tube and monitoring site region of Aberdeen for 2012. It should be noted that the regions that have been predicted to have higher standard errors have no monitoring site or diffusion tube locations situated there. Therefore, this highlights that the predicted standard errors appear to be uniformly small over the region of monitoring site and diffusion tube locations in Aberdeen throughout the months of 2012. From Figures 3.15, it can be seen again that the standard errors for the modelled data appear to be small for January and February of 2012. In Appendix C the standard errors for the rest of the months can be observed and looking at these it will highlight similar results to January and February presented in Figure 3.15.

3.3.1 Exploring the differences predicted by the modelled and observed data

Having observed the spatial surfaces produced from carrying out ordinary kriging on both the ADMS-Urban modelled data and DEFRA, monitoring site and diffusion tube data we will now investigate the annual differences in these predictions. This will help to assess how similar the annual NO_2 concentrations are being predicted. The ADMS-Urban modelled data and monitoring site and diffusion tube data will be the main focus and the differences in these predictions will be evaluated and a map of these differences will be produced. When producing these differences the main city centre of Aberdeen was focussed on and 7454 pixels in the modelled data were examined. Again these are the pixels that occupy the main city centre were the monitoring site and diffusion tubes are located. These maps are produced for both the differences of the modelled data including the roads and excluding the roads in which the background pixels are of interest and the monitoring and diffusion tube data.



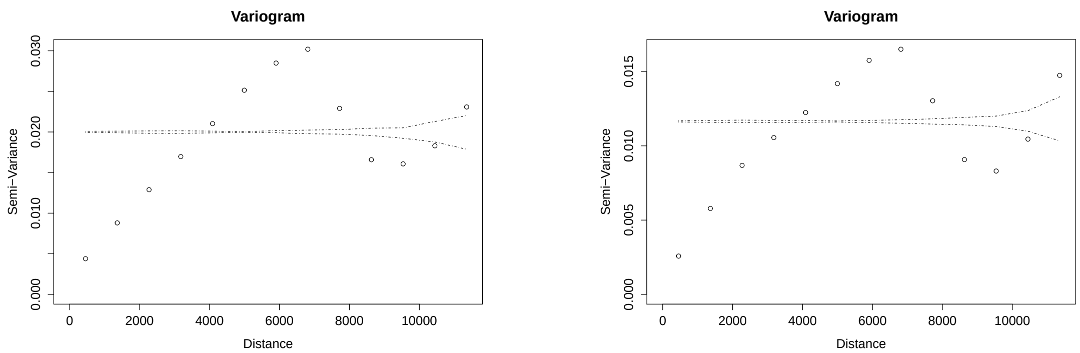
(a) Including roads

(b) Excluding roads

Figure 3.16: Map of the model measurement differences

From Figures 3.16a and 3.16b it can be observed that over the city centre region of Aberdeen the differences between both the modelled data excluding and including the roads and the monitoring and diffusion tube data are relatively small. These differences range between -0.7 μgm^{-3} and -0.1 μgm^{-3}, this indicates that the annual NO_2 modelled data is always lower than the annual NO_2 observed (monitoring site/diffusion tube) data as these differences were calculated by subtracting the observed data from the modelled data. Previously in Chapter 2 there was evidence of the modelled data being lower than the monitoring site data at Wellington Road and Anderson Drive. Observing

Figure 3.16, the annual differences appear to be smaller in the northwest of the main city centre of Aberdeen and larger in the southeast of the main city centre of Aberdeen. Figure 3.16 also suggests there is spatial structure within the roads and this can be seen especially in Figure 3.16b as the roads have been removed, however, the structure of the roads can still be detected. There appears to be spatial correlation in Figures 3.16a and 3.16b as there are areas where the pixels are producing very similar differences (almost the same). To investigate whether there is spatial correlation, variograms of the differences for both the modelled data including and excluding roads and the observed data have been produced. These variograms are highlighted below in Figure 3.17. Both Figures 3.17a and 3.17b highlight that there is strong spatial correlation present as all of the points are outside the monte carlo envelopes.



(a) Including roads

(b) Excluding roads

Figure 3.17: Monte Carlo envelopes for the variogram of the model measurement differences

3.4 Conclusion

In this chapter, through the use of different statistical techniques to explore the spatial structure in the city centre, it can be concluded that the model appears to perform reasonably well over the region of Aberdeen. It highlights that the area in Aberdeen with the highest NO_2 concentrations are in the east where the city centre of Aberdeen and Harbour are located. From the annual modelled predictions it was clear that the roads appear to have higher NO_2 concentrations and they also seem to be where the NO_2 concentrations in Aberdeen are dominate. This is highlighted as even when the roads are removed, the spatial structure of the roads can still be observed. Moverover, higher pollution levels are expected on the roads as road traffic is known as the main motivator of atmospheric pollution in Aberdeen. Moving further west it has been shown that the NO_2 concentrations appear to gradually decrease. A potential reason for this could be that you are moving away from the city centre and into the rural part of Aberdeen where the population density is smaller and there is less traffic.

Looking at the monthly modelled, DEFRA, monitoring site and diffusion tube data highlighted that the NO_2 concentrations didn't appear to change much throughout the months of the year. This may be due to the lack of data in the observed data case as there is only data for six monitoring sites and forty diffusion tube sites where only 38 diffusion tube sites have been included because of repeated locations. At these repeated locations, the monthly average concentration over these repeated locations has been considered. For both the monthly modelled and observed data, there appeared to be a region in space that was always predicted to have the highest NO_2 concentrations throughout the months of 2012 and once again this was in the city centre of the city.

Exploring the annual predicted differences between both the modelled data including and excluding the roads and the monitoring site/diffusion tube data highlighted that in both cases the differences are rather small (-0.7 μgm^{-3} to -0.1 -0.7 μgm^{-3}). Since the monitoring site/diffusion tube data were subtracted from the modelled data, this suggests that the annual modelled data is always lower than the annual monitoring site/diffusion tube data. Examining these differences also highlighted spatial structure within the roads and there also appeared to be strong spatial correlation in both cases. To explore the spatial correlation in greater detail, variograms were produced for both cases. Producing these variograms it was observed in both cases that all of the points were outside the monte carlo envelopes further indicating strong spatial correlation.

Overall, most importantly it has been shown throughout this chapter that the main area in Aberdeen that has the highest NO_2 concentrations highlighted by both the modelled and observed data is the city centre of the city. It was also highlightd by the observed data that in the northwest of the city near the airport, the NO_2 concentrations are also higher.

Chapter 4

Investigating the characteristics of the ADMS-Urban modelled pixels in space

4.1 Introduction

In this chapter functional data analysis (FDA) will be explored and explained and the process in which this analysis takes will be outlined. In particular functional principal component analysis (FPCA) and functional clustering will be investigated. The purpose of carrying out functional data analysis is to see the ways in which the ADMS-Urban pixels behave in space. By applying functional clustering to functional principal component scores, it is hoped to see which pixels are clustered together and to see if there are patterns forming in space. This analysis will only be applied to the ADMS-Urban modelled data. From exploring this analysis, it of interest to reduce the curves through FPCA and then cluster these curves using partitioning around medoids (PAM) clustering. Through clustering these curves it should help identify which pixels behave similarly. This could mean that NO_2 concentrations at certain pixels could be monitored and through doing this some idea of how the other pixels are behaving would be given.

Ramsay and Silverman (2005) state that functional data comes in various forms. These include replications which are independent, or having to work with a lone long record and may also appear as pairs of input/output variables. The aims of FDA are to illustrate the data with methods that help further analysis and present the data so as to emphasise different features. Aims of FDA also include studying vital sources of pattern and variation between the data, describing variation in a result or response variable by using input or predictor variable data and contrasting at least two sets of data regarding definite kinds of variation, where two sets of data can include various sets of repeats of identical functions, or non-identical functions for a usual set of repeats (Ramsay and Silverman, 2005).

The fundamental philosophy of FDA is to consider observed data functions as single units. The word functional when referring to observed data corresponds to the intrinsic form of the data. Functional data in practice are normally observed and recorded discretely. The data are recorded as n pairs $(t_j,\, y_j)$, and y_j is a glimpse of the function at time t_j, perhaps unclear due to error in the measurement. The continuum across which functional data are recorded is generally time. Other continua includes spatial position, frequency, weight, etc. Generally when taking functional data into consideration, a lone function x is not of interest. Instead a collection or sample of functional data is of interest and in this analysis 18319 functions are of interest. Particularly, the observation of the function x_i may be made up of n_i pairs $(t_{ij},\, y_{ij})$ where $j = 1, \ldots, n_i$. The argument values t_{ij} and interval $\mathcal{T}$ across which data are gathered may differ from record to record or may be the same (Ramsay and Silverman, 2005).

Ignaccolo *et. al* (2008) discuss in their paper, Analysis of air quality monitoring networks by functional clustering, the classification of monitoring stations by means of homogeneous clusters. Air pollutant concentrations, in particular NO_2, PM_{10} and O_3 are considered as functional data and then classification is done using functional cluster analysis, where PAM algorithm is implanted. The study region of interest is Piemonte which is in Northern Italy and this analysis is applied to the air quality monitoring network there.

An approach involving two steps was considered here. The first step involves turning discrete time series into functional data and this is done via evaluating spline coefficients. The second step of the process involves partitioning the coefficients that have been evaluated by PAM classification. Let P and i denote a general pollutant and site respectively and let the number of sites and the number of times which data are collected at be denoted by n_p and $m_{i,P}$ respectively. Ignaccolo *et. al* (2008) fit discrete data denoted by $y_{i,j}$, where $i = 1, \ldots, n_P$ and $j = 1, \ldots, m_{i,P}$ using the model

$$y_{i,j} = G_i(t_j) + \epsilon_{i,j} \tag{4.1}$$

From Equation 4.1, G_i denotes smooth functions estimated at time t_j and $\epsilon_{i,j}$ denotes the independent random errors. When turning discrete data into curves smoothing is required and to fit the curves B-spline functions are used and this will be described more later on in Section 4.2.1.

In the second step of their process to cluster objects Ignaccolo *et. al* (2008) use the non-hierarchical algorithm PAM. PAM is built on looking for k representative medoids in the dataset and groups are determined around them. PAM appears as though it takes less time as it reassigns automatically medoids in its second step called SWAP. Ignaccolo *et. al* (2008) computed the average silhouette width and this was used to choose the optimum number of clusters (Ignaccolo *et. al*, 2008). More information on this will be given later.

Abraham *et. al* (2003) state that data in various fields, for example biology, are gathered by practitioners through a procedure naturally illustrated as functional. The functional form of the data has to be taken into consideration even though data may include measurement inaccuracies as it is collected as finite vector. Abraham *et. al* (2003) put forward a clustering method of such data highlighting the functional features of the data, where the clustering procedure involves two steps. The first step involves using B-splines to fit the functional data and the second step consists of using a k-means procedure to divide the evaluated model coefficients. In this study powerful consistency of the clustering procedure is demonstrated (Abraham *et. al*, 2003).

Heimann *et. al* (2015) discuss how high spatial density and fast dependent measurements from low-cost sensor systems may make it easier to separate factors that contribute to pollutant levels that occur as a result of local emissions from those applicable to non-local sources or regional emission sources. Heimann *et. al* (2015) suggest a completely measurement-based method to obtain fundamental pollution levels from the observations. This uses the various comparative frequencies of both local and background pollution differences. In this paper it is highlighted that different factors that add to overall pollution levels can be determined if high spatial and temporal coverage of air quality observations are obtainable. The techniques carried out by Heimann *et. al* (2015) using carbon monoxide observations has extensive applicability. These consist of extra gas phase species and observations achieved using reference systems

(Heimann *et. al*, 2015).

4.2 Methodology

4.2.1 Representing functions by basis functions

The following methods explained in this section is based on Ramsay and Silverman (2005). According to Ramsay and Silverman (2005) a basis function structure is a set of functions ϕ_k which are known that are mathematically independent of one another. They also have the property that they can estimate some function arbitrarily well. This is done by taking a sum that has been weighted or linear combination of an adequately big number K of these functions. Basis functions methods indicate a function x by the following linear expansion

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{4.2}$$

with regard to K basis functions ϕ_k which are known. Equation 4.2 can also be stated in matrix notation by allowing $\mathbf{c}$ specify the vector of the coefficients c_k which is of length K and by allowing $\boldsymbol{\phi}$ be the functional vector whose components are the basis functions ϕ_k as

$$x = \mathbf{c}'\boldsymbol{\phi} = \boldsymbol{\phi}'\mathbf{c}.$$

Preferably, basis functions and the functions being approximated should have similar characteristics. This makes it more straightforward to attain a sufficient approximation using a relatively small number of basis functions represented by K (Ramsay and Silverman, 2005).

For a derivative estimate the basis selected is extremely crucial. The derivative estimate is given by

$$D\hat{x}(t) = \sum_{k}^{K} \hat{c}_k D\phi_k(t) = \hat{\mathbf{c}} D\boldsymbol{\phi}(t).$$

Slightly bad derivative estimates may be produced when bases work effectively for function estimation. One of the standards for selecting a basis may depend on if at least one of the derivatives of the approximation act sensibly. However, Ramsay and Silverman (2005) state that functional data analysis in the present day involve either a Fourier basis or a B-spline basis for periodic and non-periodic data respectively. In this analysis the data are non-periodic and therefore, a B-spline basis will be used (Ramsay and Silverman, 2005).

The spline basis system for open-ended data

For non-periodic functional data or parameters the frequent approximation structure that is selected is spline functions. When describing a spline, step one is to split the interval over which a function is to be estimated into a number of subintervals denoted by L. These L subintervals are divided by breakpoints or knots which are denoted by τ_l where $l = 1, \ldots, L - 1$. Breakpoints is the more accurate term to use. A spline is a polynomial of a defined order over every one of the intervals where order will be denoted by m. The amount of constants needed to explain the polynomial describes the order of the polynomial. The order is degree + 1 where degree is its highest power. Ramsay and Silverman (2005) said that the rule is: "*The total number of degrees of freedom in the fit equals the order of the polynomials plus the number of interior breakpoints.*" The spline regresses to being a simple polynomial when there are no interior knots (Ramsay and Silverman, 2005).

Increasing the amount of breakpoints is the central way to achieve flexibility in a spline. Across regions where the function displays the most complicated variation, more breakpoints are usually wanted. Meanwhile, less breakpoints are wanted where the function is only slightly nonlinear. Breakpoints and knots are not exactly the same thing as there can be at least two breakpoints that move at the same time to combine or be concurrent. Therefore, the amount of distinctive knot values is denoted as the breakpoint and the order of values at breakpoints where some breakpoints can be related with several knots is denoted as the knot. In the majority of applications, the knots are all definite and therefore breakpoints and knots become the same thing. A spline function is decided by the order of the polynomial segments and the knot sequence denoted by τ. To describe a spline function in the typical state of one knot per breakpoint, the amount of parameters needed is the order plus the amount of interior knots i.e $m + L - 1$ (Ramsay and Silverman, 2005).

B-spline basis for spline functions

Previously, a spline function was described, now how to build a spline function is explained. For this, a system of basis functions denoted by $\phi_k(t)$ is defined and has three essential properties. These are, every one of the basis functions $\phi_k(t)$ is a spline function itself and is given by m and τ, a spline function is any linear combination of these basis functions and any spline function described by the order m and the knot sequence τ can be conveyed as a linear combination of these basis functions. The second property holds since a multiple of a spline function, the sums and the differences of a spline function are still a spline function (Ramsay and Silverman, 2005).

At the boundaries of B-spline basis functions differentiability is lost and this makes sense because other than on the interval that the data has been gathered there is generally no information about what the function that is being evaluated is doing outwith this interval. The chance that the function may be discontinuous outwith the interval that the data has been gathered on is thus allowed for. There is m knots placed at the boundaries of B-spline basis functions to deal with this boundary behaviour. This means τ, the knot sequence, is increased at the boundaries to add an extra $m - 1$ repeats of the boundary knot value when B-splines are actually calculated (Ramsay and Silverman, 2005).

Frequently the notation $\mathrm{B}_k(t,\ \tau)$ is used. This specifies the value at point t of the B-spline basis function which is determined by the breakpoint sequence τ. The number of the biggest knot at or to the direct left of value t is given by k. In this system the $m - 1$ knots added to the first breakpoint are also included and the number of basis functions given by this notation is $m + L - 1$, as needed in the typical case where every one of the interior knots are discrete. This notation states that a spline function $\mathrm{S}(t)$ with interior knots which are discrete is given by

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau).$$

Advice on where the interior breakpoints or knots τ_l should be placed will now be discussed. Equal spacing is used as a default by the majority of applications. This is suitable given the air quality data are fairly equally spaced. If however, the data are not equally spaced it may be more sensible to place a knot at every j^{th} data point. Here j is a known number beforehand. Interior knots may also be placed at the quantiles of the argument distribution or in areas known to contain high curvature more knots placed and where there is less curvature, not as many knots placed. Data-driven techniques for where breakpoints should be placed also exist. Some of these techniques start off with a dense set of breakpoints and ones that are not needed are removed. This is done by an algorithmic method which is very alike variable selection methods used in multiple regression. For an example on this see Friedman and Silverman (1989). Another potential option would be optimising the fitting criterion respecting where the knots are placed at the same time that coefficients of the expansion are evaluated. Computational difficulties can arise here however, as fitting criteria can differ in extremely complex ways as a function of placement of the knots. De Boor (2001) explores some helpful methods for enhancing knot placement and Eubank (1999) and Green and Silverman (1994) gives a more intermediate level understandable introduction to splines (Ramsay and Silverman, 2005).

4.2.2 Choosing the number K of basis functions

Choosing K can be a difficult task and making K too big or small can both cause problems. When K is bigger the fit to the data is improved but there is also the possibility of fitting noise or variation that should be disregarded. However, when K is smaller, crucial features of the smooth function x that is being evaluated may be missed (Ramsay and Silverman, 2005).

The generalised cross-validation or GCV method

Craven and Wahba (1979) developed the generalised cross-validation measure (GCV), GCV is often used in the spline smoothing literature. The criterion is generally given by

$$\mathbf{GCV}(\lambda) = \frac{n^{-}1\mathbf{SSE}}{[n^{-}1\mathrm{trace}(\mathbf{I} \quad \mathbf{S}_{\phi,\lambda})]^2}, \tag{4.3}$$

where $\mathbf{S}_{\phi,\lambda}$ is the smoothing operator and is given by

$$\mathbf{S}_{\phi,\lambda} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda\mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}. \tag{4.4}$$

From Equation 4.4, $\mathbf{\Phi}$ is a n by K matrix that is made up of the values of the K basis functions at the n sampling points, $\mathbf{W}$ denotes the weight matrix and this allows for potential covariance structure between residuals, $\mathbf{y}$ denotes the vector of discrete data to be smoothed and $\mathbf{R}$ is given by

$$\int D^m\phi(s)D^m\phi'(s)\,ds. \tag{4.5}$$

From Equation 4.5 D^m(t) denotes the m^{th} derivative of t where t is some function. It can be more exhibiting to use the following expression for $\mathbf{GCV}(\lambda)$

$$\mathbf{GCV}(\lambda) = \Big(\frac{n}{n - df(\lambda)}\Big)\Big(\frac{\mathbf{SSE}}{n - df(\lambda)}\Big),$$

which is the same as Equation 4.3 and where df denotes the degrees of freedom and is given by

$$df(\lambda) = \text{trace}\, \mathbf{S}_{\phi,\lambda}.$$

Minimising **GCV** regarding λ will definitely involve trying various values of λ. By carrying out an initial generalized eigenanalysis, the calculation of $\mathbf{GCV}(\lambda)$ can be substantially made faster. Criterion **GCV** can be given in terms of $\mathbf{Y}$ where $\mathbf{Y}$ denotes the n by N data matrix, $\mathbf{\Phi}$ the $n \times K$ matrix containing the values of the basis function and $\mathbf{R}$ which denotes the order K penalty matrix. This gives

$$\mathbf{GCV}(\lambda) = \frac{n\,\text{trace}\{\mathbf{Y}'[\mathbf{I} - \mathbf{S}_{\phi,\lambda}]^{-2}\mathbf{Y}\}}{\{\text{trace}[\mathbf{I} - \mathbf{S}_{\phi,\lambda}]\}^2}$$

where $\mathbf{S}_{\phi,\lambda}$ the "hat" matrix is defined as

$$\mathbf{S}_{\phi,\lambda} = \mathbf{\Phi M}(\lambda)^{-1}\mathbf{\Phi}'\mathbf{W}$$

and from this

$$\mathbf{M}(\lambda) = \mathbf{\Phi}'\mathbf{W\Phi} + \lambda\mathbf{R}.$$

Every time λ is altered there is no need to invert $\mathbf{M}(\lambda)$ but a linear system of equations is required to be solved for which it is the coefficient matrix. By initially solving the generalised eigenvalue problem this can be avoided. The generalised eigenvalue problem is given by

$$\mathbf{RV} = \mathbf{\Phi}'\mathbf{W\Phi VD}. \tag{4.6}$$

From Equation 4.6, the matrix of eigenvalues of $\mathbf{R}$ is denoted by $\mathbf{D}$ in the metric determined by **Φ'WΦ** and $\mathbf{V}$, where the columns of $\mathbf{V}$ correspond to the eigenvectors of $\mathbf{R}$, fulfil the orthogonality condition

$$\mathbf{V}'\mathbf{\Phi}'\mathbf{W\Phi V} = \mathbf{I}.$$

If **Φ'WΦ** is nonsingular the generalised eigenvalue problem has a solution and this is the only case it has one. If knots are positioned at every data point **Φ'WΦ** will not be nonsingular and therefore there is no solution to the generalised eigenvalue problem.

Checking $\mathbf{\Phi}$'$\mathbf{W\Phi}$ for singularity is always recommended (Ramsay and Silverman, 2005).

For any new value of λ, the essential inverse can be defined in a competent way as

$$\mathbf{M}(\lambda)^{-1} = \mathbf{V}(\mathbf{I} + \lambda\mathbf{D})^{-1}\mathbf{V}'. \tag{4.7}$$

Equation 4.7 can be written in the following way as the matrix now being inverted is diagonal. Additionally computing the derivative of $\mathbf{GCV}(\lambda)$ requires computing the matrix

$$\mathbf{M}(\lambda)^{-1}\mathbf{\Phi}'\mathbf{W\Phi M}(\lambda)^{-1} = \mathbf{V}(\mathbf{I} + \lambda\mathbf{D})^{-2}\mathbf{V}'$$

so that supplying a derivative value to a numerical optimisation algorithm also helps in terms of computation and the amount of evaluations of $\mathbf{GCV}(\lambda)$ will probably be reduced greatly (Ramsay and Silverman, 2005).

4.2.3 Functional Principal Component Analysis

It is suggested by Ramsay and Silverman (2005) that a main approach to consider is principal components analysis (PCA) of functional data. A principal component analysis supply's a way of studying covariance structure that can be much more instructive and can enhance, or even replace entirely, a straight inspection of the variance-covariance function (Ramsay and Silverman, 2005).

PCA for multivariate data

Firstly, PCA for multivariate data will now be explained in this section before going to describe the ways in which this adapts and changes when looking at functional data. The main idea of taking a linear combination of variable values is utilised continuously in multivariate statistics. For example,

$$f_i = \sum_{j=1}^{p} \beta_j x_{ij},\ i = 1, \ldots, N. \tag{4.8}$$

From Equation 4.8 β_j denotes a weighting coefficient applied to x_{ij} of the j^{th} variable where x_{ij} represents the observed values. Weights are chosen in the multivariate case so as to emphasise or present types of variation that are clearly illustrated in the data. The method for explaining principal components analysis, which determines sets of normalised weights that give the maximum variation in the f_i's can be given as follows.

The first step involves finding the weight vector which is denoted by $\boldsymbol{\xi}_1 = (\xi_{11}, \ldots, \xi_{p1})'$ for which the values of the linear combination given by

$$f_{i1} = \sum_j \xi_{j1} x_{ij} = \boldsymbol{\xi}_1' x_i$$

have the biggest attainable mean square given by $N^{-1} \sum_i f_{i1}^2$ conditional on the following constraint

$$\sum_j \xi_{j1}^2 = ||\boldsymbol{\xi}_1||^2 = 1.$$

The second step and every step there after are carried out and this may be done up to a maximum of p steps which denotes the number of variables. On the mth step a new weight vector denoted by $\boldsymbol{\xi}_m$ is calculated. This new weight vector has elements ξ_{jm} and new values $f_{im} = \boldsymbol{\xi}_m' x_i$ where these new values now have maximum mean square, conditional on the constraint $||\boldsymbol{\xi}_m||^2 = 1$ and the $m - 1$ extra constraints where there may be 1 or more

$$\sum_j \xi_{jk} \xi_{jm} = \boldsymbol{\xi}_k' \boldsymbol{\xi}_m = 0,\ k < m.$$

The motive for carrying out the first step of the principal components analysis method is that by finding the maximum of the mean square, the most powerful and significant mode of variation in the variables are determined. For the problem to be clearly described, the constraint of the unit sum of squares on the weights is needed. The most significant modes of variation are found again on the second step of the procedure and every step there after but this time so that they specify something different, it is essential that the weights defining them are orthogonal to the weights determined already (Ramsay and Silverman, 2005).

Principal component scores are denoted by the values of the linear combinations f_{im}. These are usually of great assistance in explaining what these significant components of variation indicate with regards to the features of particular cases or replicates (Ramsay and Silverman, 2005).

Defining PCA for functional data

In a functional context the equivalent of variable values are function values $x_i(s)$. This means that the discrete index j used previously in the multivariate case has been replaced by the continuous index s. When vectors are being considered, it was suitable to compute the inner product as a way of joining a weight vector β with a data vector x, where the inner product is given by

$$\beta' x = \sum_j \beta_j x_j.$$

If the weight vector and data vector are functions say $\beta(s)$ and $x(s)$ respectively, summations over j are replaced by integrations over s and the inner product is now given by

$$\int \beta x = \int \beta(s)x(s)ds. \tag{4.9}$$

Within the principal components analysis, the weights are now functions and have values $\beta_j(s)$. Using the same notation as used in Equation 4.9, the principal component scores corresponding to weight β are defined as follows

$$f_i = \int \beta x_i = \int \beta(s)x_i(s)ds.$$

The first step of the functional PCA involves choosing the weight function denoted by $\xi_1(s)$ to maximize $N^{-1}\sum_i f_{i1}^2 = N^{-1}\sum_i(\int \xi_1 x_i)^2$ conditional on the continuous analogue $\int \xi_1(s)^2 ds = 1$ of the unit sum of squares constraint. To indicate the squared norm $\int \xi_1(s)^2 ds = \int \xi_1^2$ of the function ξ_1 the notation $||\xi_1||^2$ is utilised. The cumulative variance is used to choose the number of components (Ramsay and Silverman, 2005).

4.2.4 Clustering

K-means (MacQueen, 1967; Hartigan and Wong, 1979) is a widely chosen partitioning procedure used to date. It is an iterative procedure that for a stated number of clusters reduces the within-class sum of squares (MacQueen, 1967; Hartigan and Wong, 1979). The first step of the procedure is to initially approximate the cluster centres. Then each observation is situated into the nearest cluster. The next stage is to update the cluster centres and the full procedure is iterated until convergence (Charrad *et al.*, 2014). In this thesis partitioning around medoids clustering is used and this method will now be described.

Partitioning around Mediods Clustering

Partitioning around medoids clustering was used and the clusters were chosen using the optimum average silhouette width. Charrad *et al.* (2014) state that the silhouette index (SI) is defined as

$$\mathrm{SI} = \frac{\sum_{i=1}^{n} S(i)}{n}, \ \mathrm{SI} \in [-1, 1],$$

where

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}},$$

and $a(i)$ denotes the average dissimilarity of the i^{th} object to all other objects of cluster C_r and $b(i)$ denotes the minimum value of the average dissimilarity of the i^{th} object to all objects of cluster C_s where r $\neq$ s. When deciding on the optimal number of clusters that should be chosen, the maximum value of the index is used (Charrad *et al.* 2007).

Partitioning Around Medoids (PAM) was a clustering algorithm suggested by Kaufman and Rousseeuw (1990). This clustering algorithm maps a distance matrix into a number of clusters which have been defined. PAM has an extremely nice feature in that it allows clustering regarding any described distance metric. Furthermore, the medoids are robust portrayals of the cluster centers. In the usual case that numerous elements do not belong nicely to any cluster, this is especially vital (van der Laan *et. al*, 2002).

As mentioned in the Ignaccolo *et. al* (2008) paper discussed in the introduction of this chapter, PAM is made up of two stages namely BUILD and SWAP. Firstly, an initial clustering is achieved by the consecutive selection of representative objects. This is done until k objects have been established. The first object has the smallest sum of dissimilarities to all other objects and is situated nearest the middle in the set of objects. At each step thereafter another object is chosen and this object reduces the objective function as much as it can. The following steps are performed in order to find this object:

1. At first an object that has not yet been chosen say i is taken into consideration.

2. Then take into account an object j which hasn't yet been chosen and compute the difference between its dissimilarity which is denoted by D_j with the object it is most similar too that has been chosen already and $d(j, i)$ which denotes its dissimilarity with object i.

3. Object j will contribute to the option of choosing object i when this difference is non-negative and $C_{ij} = \max(D_j - d(j, i), 0)$ is computed.

4. The total gain achieved is computed by choosing object i $\sum_j C_{ji}$.

5. Select the object i that has not been chosen yet which

$$\text{maximises}_i \sum_j C_{ji}.$$

This procedure is maintained until k objects have been established. Secondly in the SWAP stage of the PAM algorithm, the aim is to enhance the set of representative objects and hence also to enhance all the clustering produced by this set. To do this all pairs of objects (i, h) are considered where object i and h have been chosen and not chosen respectively. When a swap is performed it is decided what effect is achieved on the value of the clustering. In order to compute this effect between i and h steps 1 and 2 of the following computations are performed:

1. Firstly take into account an object j which hasn't yet been chosen and compute its input denoted by C_{jih} into the swap. To do this the following is carried out:
   (a) C_{jih} is zero if one of the other representative objects is closer than both i and h to j.
   (b) Two cases has to be considered if j is closer to i than to any other chosen representative object. These cases are as follows:
      i. If j is nearer to h than to the representative object that is second nearest then $d(j, h) < E_j$ where E_j represents the dissimilarity between j and the representative object that is second most similar. Here the input into the swap of object j between object i and object h is given by $C_{jih} = d(j, h) - d(j, i)$.
      ii. If the case is that j is at least further away from h than from the representative object that is second nearest then $d(j, h) \geq E_j$. Here the input into the swap of object j between object i and object h is given by $C_{jih} = E_j - D_j$.
   (c) If at minimum one of the other representative objects is closer than i to j but j is nearer to h than to any other representative object then the input into the swap of object j between object i and object h is given by $C_{jih} = d(j, h) - D_j$.
2. Then the overall outcome of a swap is computed by adding the inputs C_{jih}.

In the following two steps it is determined whether to perform a swap.

3. Choose the pair of objects i and h that

$$\underset{i,h}{\text{minimises}} \sum_j C_{jih}.$$

4. Depending on whether the minimum $\sum_j C_{jih}$ is negative, positive or 0 depends on whether a swap should be carried out. If the result is positive the swap is performed and the algorithm returns to step 1, however if the result is positive or 0, by performing a swap the value of the objective cannot be reduced and the algorithm ends (Kaufman and Rousseeuw, 1990).

Partitioning around medoids clustering will be carried out on the principal component scores produced from carrying out functional PCA on the ADMS-Urban modelled pixels. This is done in order to investigate the way in which the modelled pixels are clustered together and to observe if these clusters form patterns in space and this will give insight into if there are areas of Aberdeen that have higher NO_2 concentrations.

4.3 Results

Throughout this analysis the main packages in *R* used were *fda* (CRAN, 2014), *fda.usc* (CRAN, 2015b), *fpc* (CRAN, 2015c), *cluster* (CRAN, 2015d) and *ggplot2* (CRAN, 2015e). When FPCA and clustering were carried out, the full region of the ADMS-Urban modelled pixels were not considered but instead it was of interest to consider six cases. These included zooming into the city centre region where the monitoring sites and most of the diffusion tube locations were situated which consisted of 7454 pixels including roads, looking at 7454 background pixels, and looking at these again but this time for summer and winter months individually. However, when considering summer and winter background pixels, all 10201 pixels were easily able to be investigated as these only contained data for three months each. Looking at the background pixels would highlight the impact roads were having on the clustering and investigating summer and winter months where June, July and August represent summer and January, February and March represent winter is of interest to see if clusters change depending on the weather. The main reason for not taking the full ADMS-Urban region into consideration was again due to the computational challenges. To estimate the smooth functions, b-spline basis was used as the data were non-periodic and then functional principal components analysis was applied to the smoothed functions. Generalised cross-validation was used to choose the number of basis functions and these were computed using the *min.basis* function in *R* within the *fda.usc* package (CRAN, 2015b).

The number of components were chosen using the cumulative variance explained and the number of clusters were chosen using the optimum average silhouette width. An example plot of the cumulative variance explained and summary of the optimum average silhouette width are given below:

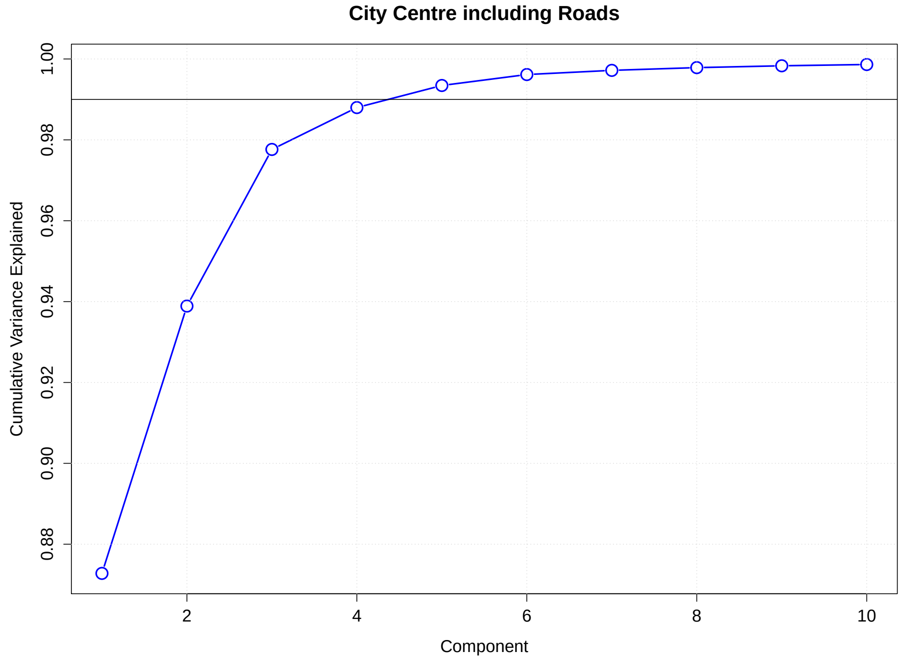
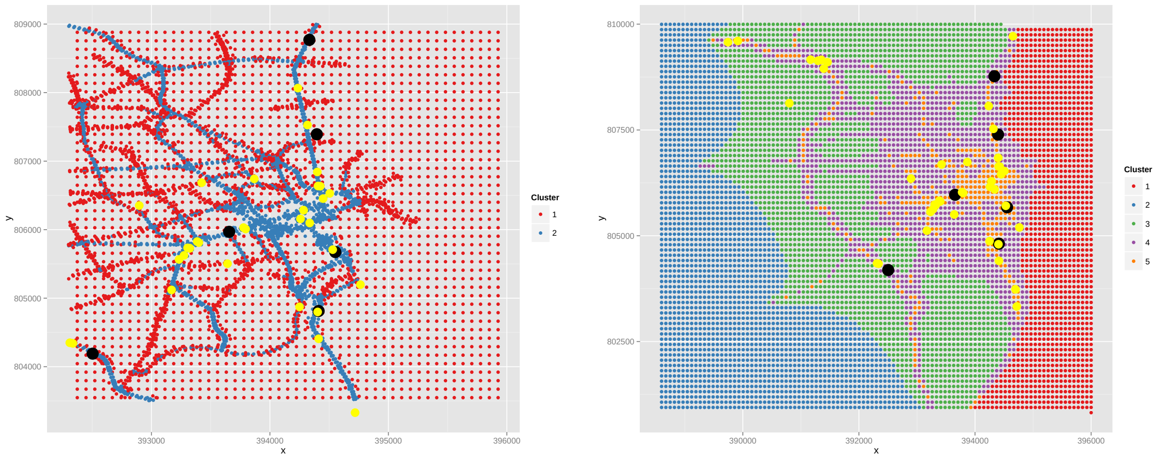


Figure 4.1: Cumulative Variance Explained against Component for the zoomed in city centre region of 7454 pixels

Table 4.1: Summary of the optimum average silhouette width for the zoomed in city centre including roads region of 7454 pixels

Number of clusters	Average silhouette width
2	0.460
3	0.333
4	0.366
5	0.315
6	0.262
7	0.258
8	0.256
9	0.260
10	0.267

From Figure 4.1 the horizontal line represents that 99% of the cumulative variance has been explained and this was used to decide on the number of components to use. Hence, looking at the plot it was suggested that 4 components should be chosen and the rest of the components were chosen in a similar manner. Table 4.1 highlights the average silhouette widths for the number of clusters ranging from 2 to 10, observing this table we see that the largest average silhouette width occurs when the number of clusters is 2. This suggests that 2 clusters should be chosen when carrying out clustering in the zoomed in city centre region of Aberdeen including the roads. Similarly the rest of the clusters were chosen using the exact same method. After the number of components and clusters were decided for all six cases the following results were produced:

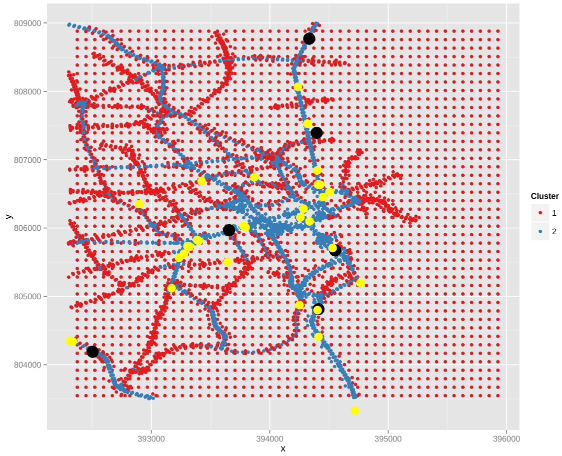


(a) Map of Aberdeen city centre

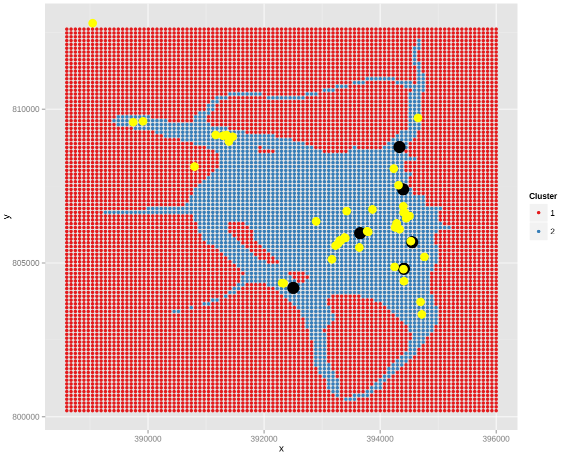
(b) Map of background pixels

Figure 4.2: Maps of Aberdeen highlighting the clusters using the partitioning around medoids clustering algorithm (Black circles represent the 6 monitoring site locations in Aberdeen and the yellow circles represent the diffusion tube locations)

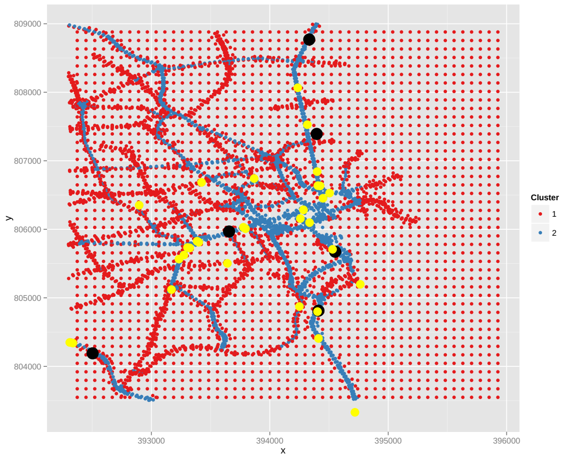
Figure 4.2b represents the background pixels where the road pixels have been removed to see if the roads were having an effect on the clustering allocations. From Figure 4.2a it appears as though the higher NO_2 concentrations mostly those on the roads have been clustered together and the background concentrations and roads with lower NO_2 concentration have been clustered together. Removing the roads and running the analysis again on just the background pixels shows that the NO_2 concentrations have again been clustered by how high and low they are. The cluster highlighted in purple almost appears to outline the roads in Aberdeen and then the cluster represented by orange appears to highlight the even higher NO_2 concentrations and this appears to be where most of the monitoring sites and diffusion tubes sit.



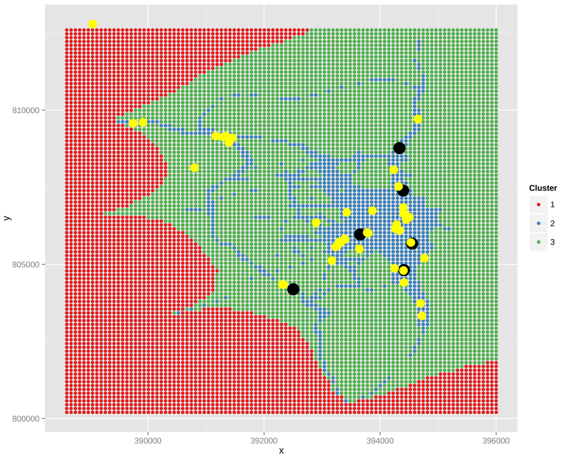
(a) Map of Aberdeen city centre in the summer



(b) Map of background pixels in the summer



(c) Map of Aberdeen city centre in the winter



(d) Map of background pixels in the winter

Figure 4.3: Maps of Aberdeen highlighting the clusters in the summer (June, July and August) months and winter (January, February and March) months using the partitioning around medoids clustering algorithm (Black circles represent the 6 monitoring site locations in Aberdeen and the yellow circles represent the diffusion tube locations)

Figures 4.3a and 4.3c emphasise once again that it seems as though the higher NO_2 concentrations mostly those on the roads have been clustered together and the background concentrations and roads with lower NO_2 concentration have been clustered together. In fact, there doesn't appear to be much of a change when looking at summer and winters months individually and looking at the city centre across the full year as Figures 4.3a and 4.3c and 4.2a are looking to be very similarly clustered. Once more in Figures 4.3b and 4.3d it can be observed that removing the roads and running the analysis again on just the background pixels shows that the NO_2 concentrations have again been clustered by how high and low they are. The blue cluster in the both the right hand plots appears to outline the roads in Aberdeen and these are where most of the monitoring sites and diffusion tubes sit.

The cluster mean curves where produced for the clusters represented in Figures 4.2 and 4.3. The daily mean NO_2 concentration data for each monitoring site has also been produced and this is to assist in examining how close the cluster mean curves and the daily mean NO_2 concentration data are at each site. The results are given in Figures 4.4, 4.5, 4.6, 4.7, 4.8 and 4.9. In Figure 4.5, clusters 4 and 5 from Figure 4.2b have been shown as these clusters are where the monitoring sites and diffusion tubes were located.

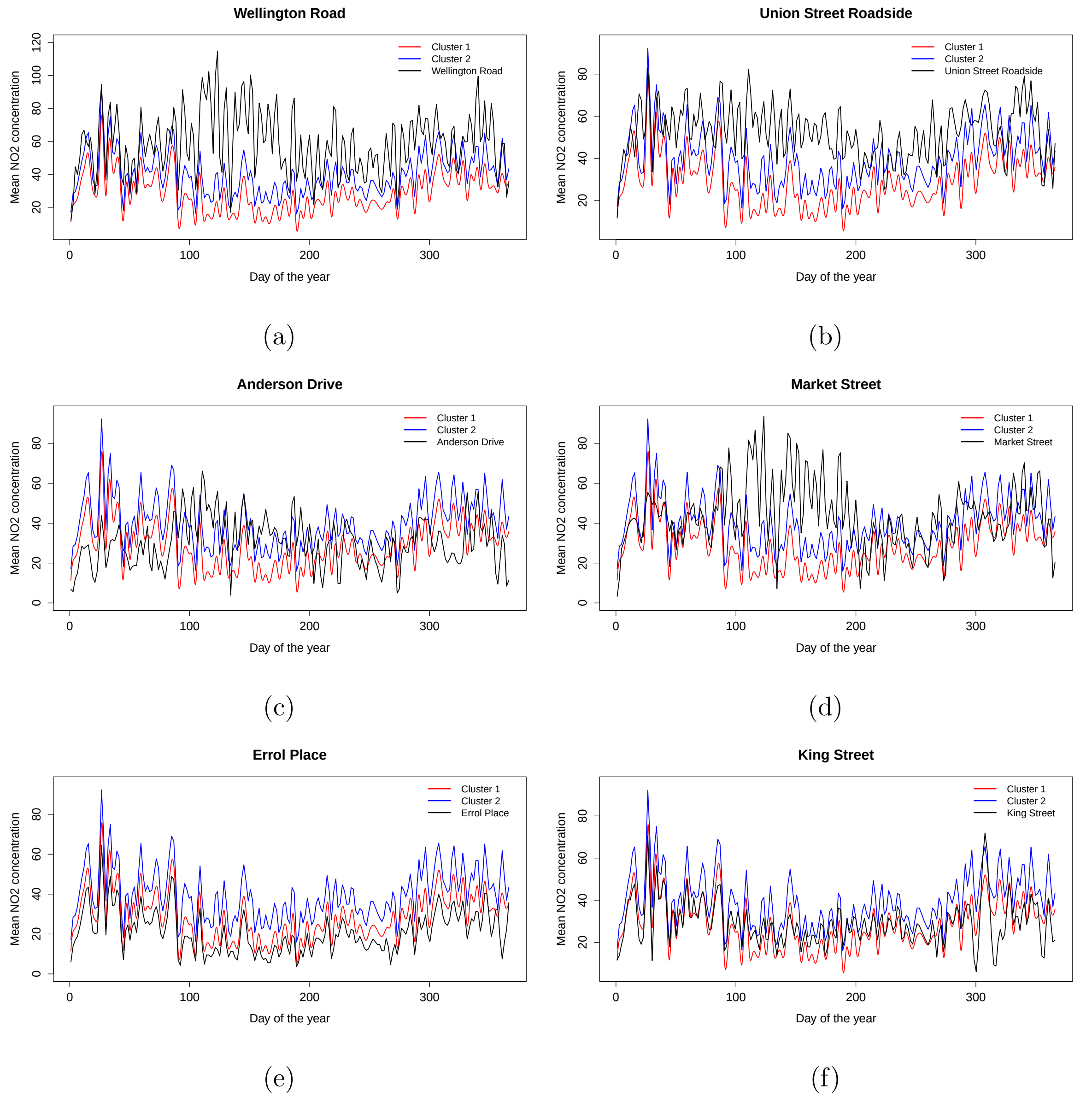


Figure 4.4: Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the main city centre of Aberdeen including the roads and monitoring daily NO_2 data for each site represented by the black line

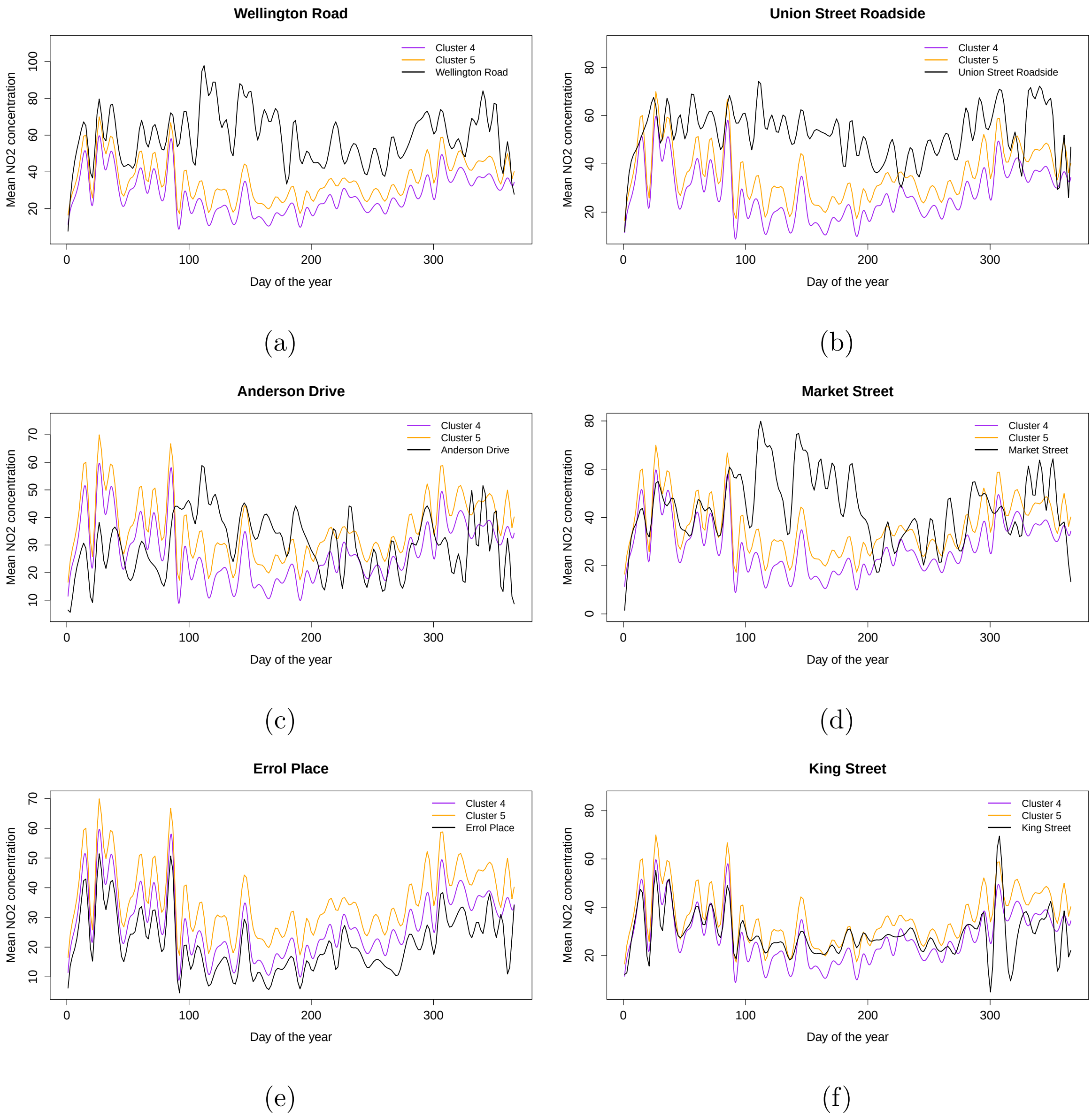


Figure 4.5: Cluster mean curves for cluster 4 and cluster 5 represented by the purple and orange line respectively in the background pixels of Aberdeen and monitoring daily NO_2 data for each site represented by the black line

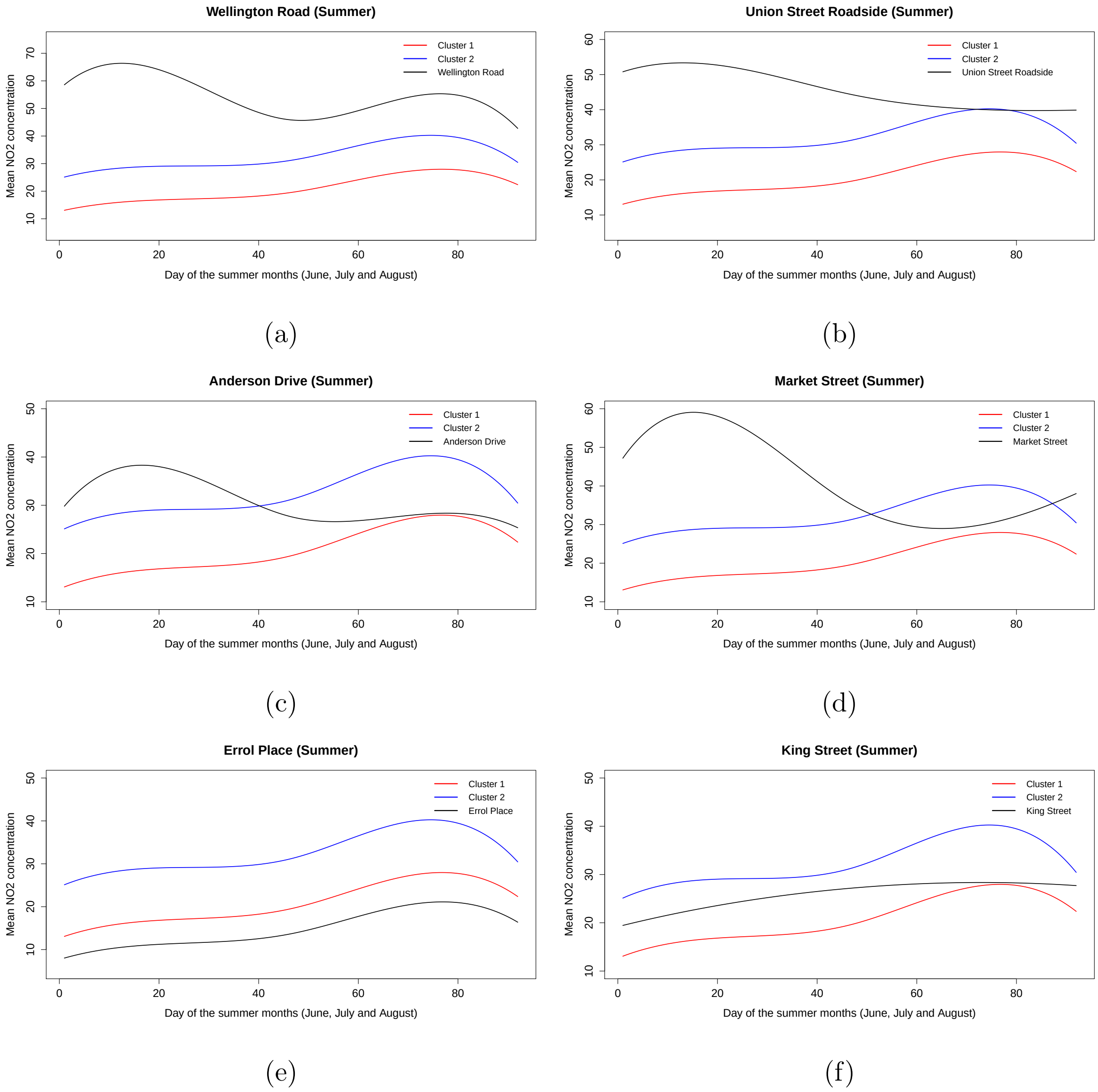
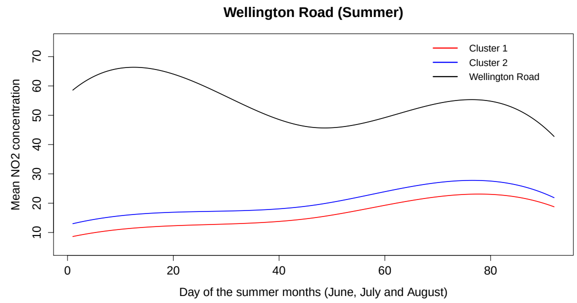
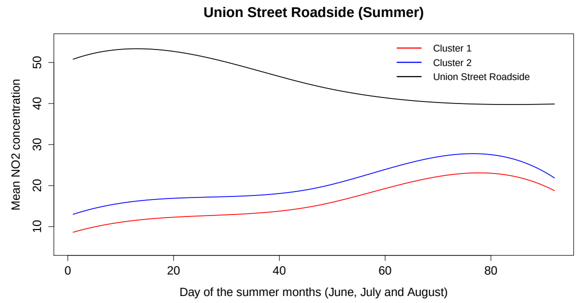


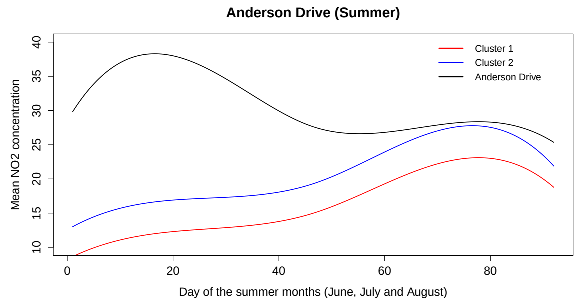
Figure 4.6: Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the main city centre of Aberdeen in the summer months (June, July and August) including the roads and monitoring daily NO_2 data for each site represented by the black line



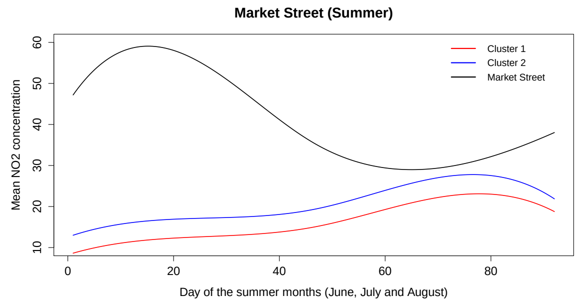
(a)



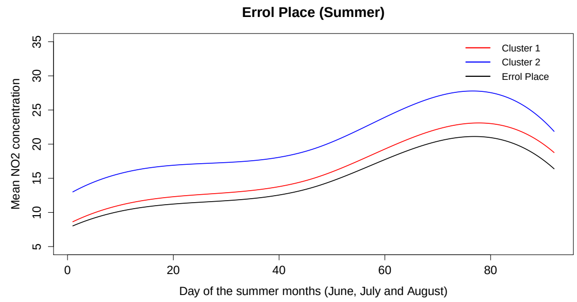
(b)



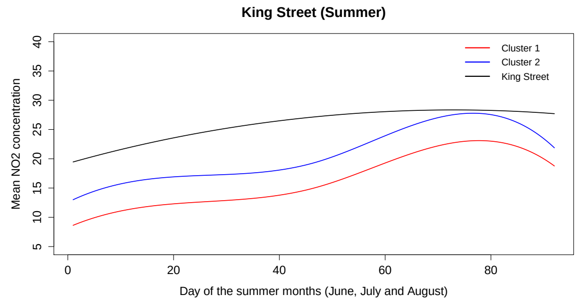
(c)



(d)

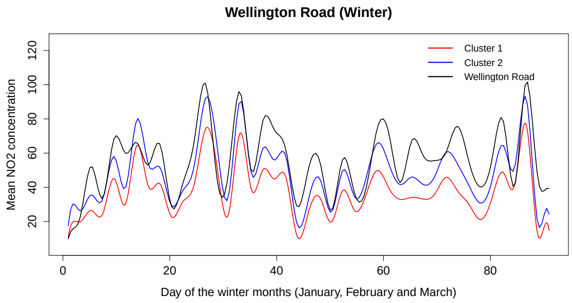


(e)

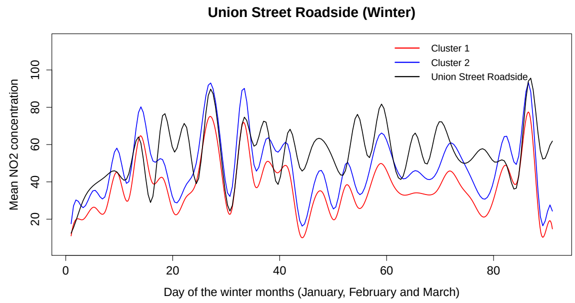


(f)

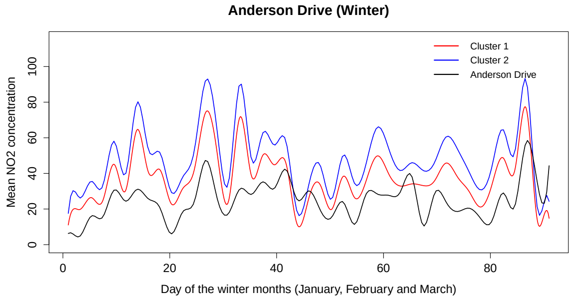
Figure 4.7: Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the background pixels of Aberdeen in the summer months (June, July and August) and monitoring daily NO_2 data for each site represented by the black line



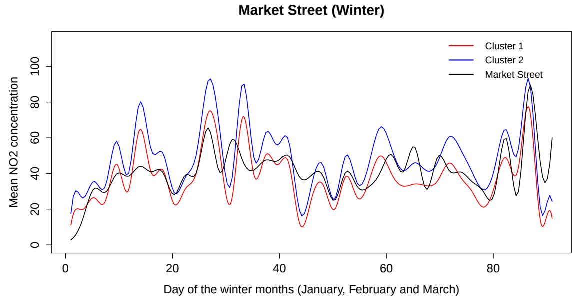
(a)



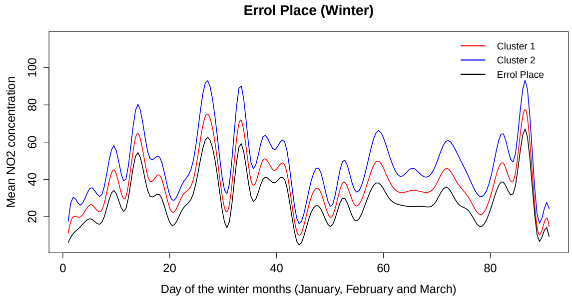
(b)



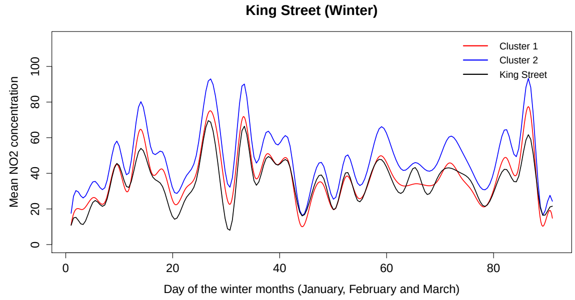
(c)



(d)

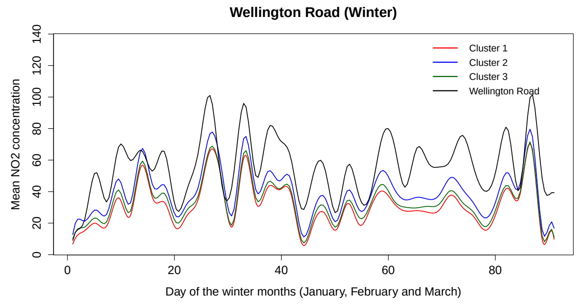


(e)

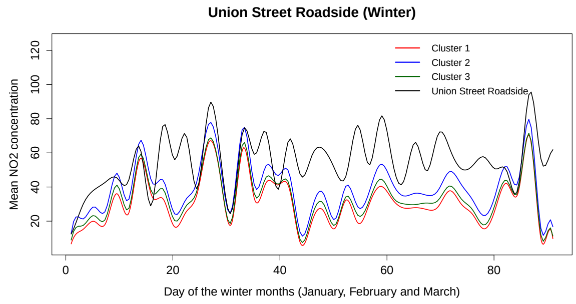


(f)

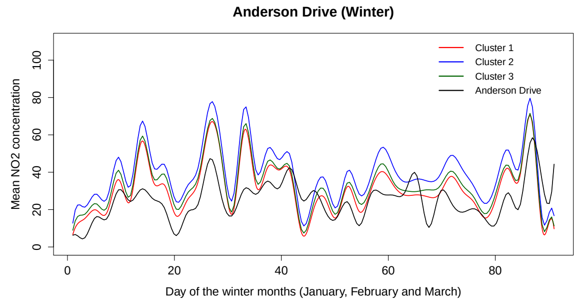
Figure 4.8: Cluster mean curves for cluster 1 and cluster 2 represented by the red and blue line respectively in the main city centre of Aberdeen in the winter months (January, February and March) including the roads and monitoring daily NO_2 data for each site represented by the black line



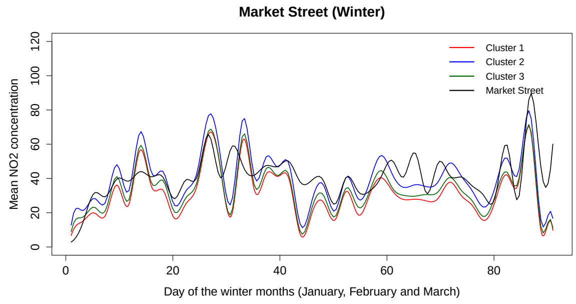
(a)



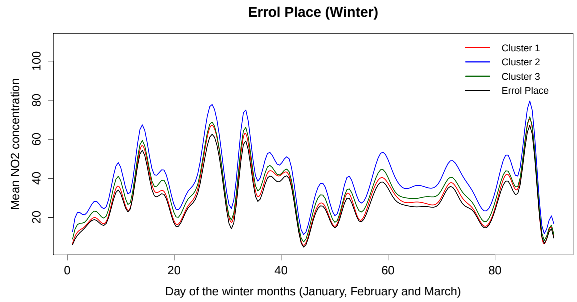
(b)



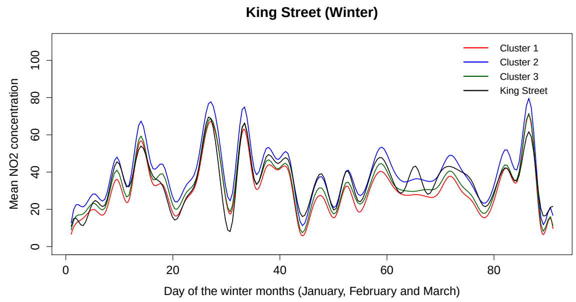
(c)



(d)



(e)



(f)

Figure 4.9: Cluster mean curves for cluster 1, cluster 2 and cluster 3 represented by the red, blue and green line respectively in the background pixels of Aberdeen in the winter months (January, February and March) and monitoring daily NO_2 data for each site represented by the black line

As previously mentioned Figure 4.2b emphasised that the monitoring site locations appeared to be located around cluster 4 and cluster 5 and hence in Figure 4.5 it was of interest to compare the monitoring data with these clusters to see how similar they were. From both Figures 4.4 and 4.5 it is highlighted that between day 100 (9^{th} of April) and day 200 (18^{th} of July), the cluster mean curves and the monitoring data at monitoring sites Wellington Road, Union Street Roadside and Market Street appear not to be well calibrated with the monitoring data observing higher NO_2 concentrations. It can also be suggested that the monitoring sites Errol Place and King Street appear to follow the cluster mean curves more accurately with the monitoring site Wellington Road once again behaving most poorly. However, it should be noted once again that the monitoring data from Errol Place was used in the model set-up and therefore would be expected to perform better. Observing Figures 4.6, 4.7, 4.8 and 4.9 highlights that the monitoring sites appear to follow the cluster mean curves more accurately in the winter months (January, February and March) than the summer months (June, July and August). Previously in Chapter 2, it was concluded in POT analysis that the number of exceedances for both the modelled and monitoring data was more accurate in the winter months compared with summer. It should also noted that again generalised cross-validation was used to choose the number of basis functions in all cases and these were computed using the *min.basis* function in *R* within the *fda.usc* package (CRAN, 2015b).

4.4 Conclusion

From this chapter, it can be concluded that by applying FPCA to the smoothed functions and applying functional clustering to the principal component scores, there appears to be a dominance of the roads in the clusters. This can be seen for the main city centre of Aberdeen over the full year 2012 and also for the summer and winter months of 2012. Even once these roads have been removed and the background pixels are investigated, there still appears to be a pattern forming of the roads appearing as a cluster. This emphasises that the roads are the main cause of concern in Aberdeen in terms of air quality and that road traffic is a major issue in Aberdeen.

Through calculating cluster mean curves, it was shown that once more Wellington Road performed most poorly as the daily mean monitoring data was much larger than the cluster mean curves. However, the daily mean monitoring data at Errol Place and King Street appeared to follow the cluster mean curves more accurately. This reinforces what has been said throughout this thesis that the modelled and monitoring data do not appear to be well calibrated at Wellington Road. Also computing the cluster mean curves for the winter (January, February and March) and summer (June, July and August) months emphasised that in the winter the monitoring site data followed the cluster mean curves more precisely.

Overall, in this chapter through carrying out clustering on the modelled pixels it can be concluded that the roads dominant the clusters in Aberdeen. Furthermore, by computing cluster mean curves, it is found that the daily mean monitoring NO_2 data is much higher than the cluster mean curves at Wellington Road and that monitoring data is better calibrated with the cluster mean curves in the winter than the summer.

Chapter 5

Functional calibration of the ADMS-Urban model output

5.1 Introduction

The aim of carrying out Functional Regression is to explore in a functional context how well the monitoring and modelled data are calibrated and also to investigate how well the diffusion tube and modelled data are calibrated. By carrying out this piece of analysis, it was hoped to produce an overall slope for the six monitoring sites and forty diffusion tubes and this would assist us in knowing overall how well the observed and modelled data are calibrated. Previously in Chapter 2, deming regression was carried out and a slope estimate for each site was produced separately, where as here an overall slope function for all six sites will be produced. It is beneficial to investigate and produce this as it is a way of summarising and bringing together how good the model predictions are at these six monitoring sites throughout the year 2012.

To do this, the same procedure is carried out that was carried out for the clustering case in the sense that the smooth functions are estimated for both the modelled and observed data using a b-spline basis. Following this a functional linear model is then fit to the data, where the smoothed functions for the ADMS-Urban modelled data are the response variable and the smoothed functions for the observed data are the explanatory variable. The response y and explanatory variable x are both functions of t where t corresponds to time and denotes the days of the year. It should be noted that models for the monitoring data and diffusion tube data will be carried out separately but will both be carried out in a very similar manner. The effect is concurrent which highlights that x only effects $y(t)$ via its value $x(t)$ at day of the year t. This analysis is suitable to compare observations time point by time point (Ramsay and Silverman, 2005).

Yen *et. al* (2015) explore in their paper different approaches for regression where the response variable is a functional object. These approaches are investigated using the *R* package *FREE*, where this package concentrates on simple application and interpretation of function regression analyses. Machine learning and various Bayesian methods are many of the computational procedures that are implemented and these procedures are compared by Yen *et. al* 2015 using both simulated and real data. Throughout this analysis many of the procedures appeared to perform just as good as one another for the same dataset. Additionally, through carrying out this analysis, Yen *et. al* 2015 discovered that through using functional regression, functional data can be modelled directly (Yen *et. al*, 2015).

In their paper Faraway (1997) consider functional responses when measurements are recorded over time. It is expressed by Faraway 1997 that if there is a smooth functional response $y(t)$, explanatory variables x which are known and parameter functions represented by $\beta(t)$ which have to be evaluated then functional regression analysis relates $y(t)$ to x by a linear combination of $\beta(t)$. The model fitted by Faraway (1997) in this analysis takes the typical form

$$y(t) = x^T\beta(t) + \epsilon(t),$$

where $y(t)$ denotes a vector of response functions, $\beta(t)$ denotes a vector of functions and X is the well known design matrix with dimensions $n \times p$, composed from the p-vector valued explanatory variables denoted by x_i, $i = 1, \ldots, n$. Furthermore, $\epsilon(t)$ represents a vector of error functions. The method undertaken in this study presents a different approach to use in biological sciences instead of using the usual longitudinal data approaches (Faraway, 1997). Cuevas *et. al* (2002) discuss in their paper the issue of simple linear regression when both the response and explanatory variables are functional and the design of the experiment is fixed. An estimator is put forward by Cuevas *et. al* (2002) for the fundamental linear operator and they intend to make sure the design is adequately providing useful information by proving its stability under some conditions. The classical calibration difficulty, sometimes referred to as the inverse regression is examined by Cuevas *et. al* (2002) and they examine a stable estimator

(Cuevas *et. al*, 2002).

It is stated in Ramsay *et. al* (2009) that the concurrent functional linear model is very similar to the varying-coefficients model. Hastie and Tibshirani (1993) discuss in their paper a group of regression and generalised regression models. Within these models the coefficients are able to differ as smooth functions of other variables. They suggest general methods for evaluating the flexibility of the models and explain how this group of models joins together two sets of models namely generalised additive models and dynamic generalised linear models into one general structure. In their paper, Hastie and Tibshirani (1993), apply their analysis to the proportional hazards model for survival data and highlight that this method supplies an alternative way of modelling deviations from the proportional hazards assumption (Hastie and Tibshirani, 1993).

In their paper Fan *et. al* (2003) explore varying-coefficient linear models. Fan *et. al* (2003) discuss that it has been a general practice to suppose that the varying coefficients are functions of a variable that is known. This variable is frequently referred to as an index. Within this paper, a group of varying-coefficient linear models in which the index is unspecified is examined. The main reason for this is to extend the modelling ability significantly and they approximate the index as a linear combination of regressors and/or other variables. Fan *et. al* (2003) then search for the index, where the search is put into effect through a hybrid backfitting method that has been newly suggested, such that the varying-coefficient model that has been obtained gives the least squares estimate to the fundamental multidimensional regression function that is not known. The t-statistic and the Akaike information criterion (AIC) are both used to chose the locally significant variables. The method is further enlarged for models with two indices and the simulation carried out in this paper suggests that the methodology Fan *et. al* (2003) put forward has considerable flexibility to model complex multivariate non linear structure. It also highlights that in practice it is reasonable with a standard current computer. The approaches presented throughout this paper are further emphasised via two examples namely the Canadian mink-muskrat data and the pound-dollar exchange rates data (Fan *et. al*, 2003).

5.2 Methodology

Within functional regression, there are different scenarios in terms of the way the models are fitted. There are three scenarios, these are:

1. The response is functional and the explanatory variable is scalar where a model of the following form

$$y_i(t) = \beta_0(t) + \sum x_{ij}\beta_j(t) + \epsilon_i(t) \tag{5.1}$$

   is fitted.

2. The response is scalar and the explanatory variable is functional where a model of the following form

$$y_i = \beta_0 + \int x_i\beta(t)dt + \epsilon_i \tag{5.2}$$

   is fitted.

3. The response and explanatory variable are both functional where there are two cases. The first case is where the effect is concurrent as mentioned previously and a model of the following form

$$y_i(t) = \beta_0(t) + \sum x_{ij}(t)\beta_j(t) + \epsilon_i(t) \tag{5.3}$$

   is fitted. The second case is more general and takes the concurrent model further. The model of the following form

$$y_i(t) = \beta_0(t) + \int_{\Omega_t} \beta_1(t,s)x_i(s)ds + \epsilon_i(t) \tag{5.4}$$

   is fitted where in this case the span or the continnum of $x_i(s)$ and $y_i(t)$ do not need to be the same (Ramsay *et. al*, 2009).

From Equations 5.1, 5.2, 5.3 and 5.4, y denotes the response and x denotes the explanatory variables. Also from these equations β_0 and β represent the intercept and slope respectively where they are a function in Equations 5.1, 5.3 and 5.4 and a scalar in Equation 5.2. The random error element is denoted by ϵ in the above equations and again is a function in Equations 5.1, 5.3 and 5.4 and a scalar in Equation 5.2. For this analysis, the model of the form in Equation 5.3 has been chosen because it is of interest to compare the ADMS-Urban modelled data and monitoring data/diffusion tube data time point by time point on the same time scale.

5.2.1 Estimation for the Concurrent Model

The model fitted is of the following form:

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{q-1} x_{ij}(t)\beta_j(t) + \epsilon_i(t). \tag{5.5}$$

From Equation 5.5, a functional observation is represented by $x_{ij}(t)$, although this could be a scalar observation or a categorical indicator. If this was the case then $x_{ij}(t)$ would be clearly thought of as a function that is uniform over time. The intercept function is represented by $\beta_0(t)$ (Ramsay *et. al* 2009). In this case, $x_{ij}(t)$ denotes the smoothed functions for the monitoring site/diffusion tube data and $y_i(t)$ denotes the smoothed functions for the ADMS-Urban modelled data where again the ADMS-Urban modelled pixels closest to the monitoring site/diffusion tube locations where used. From Equation 5.5, $\beta_j(t)$ represents the slope function and this will be the main interest throughout this analysis.

Ramsay *et. al* (2009) state that just like ordinary regression, multicollinearity must be considered among the intercept and the functional covariates. A list of difficulties are created due to multicollinearity, these include imprecision in estimates because of the error when rounding up/down, trouble in determining which explanatory variables are significant in explaining the response variable, and the lack of stability in regression coefficient estimates because of compromises between explanatory variables in explaining the variability in the response variable. A closer look at how the functional regression coefficients represented by β_j are evaluated by the function $fRegress$, found in the R package fda (CRAN, 2014), is investigated to help to comprehend the multicollinearity issue. This is done by decreasing the issue down to the solution of a set of linear equations. Let $\mathbf{Z}$ represent the functional matrix which contains these x_{ij} with dimensions $N \times q$, let β represent the vector of coefficient functions which contains every one of the regression functions with length q and let $\mathbf{y}$ denote a functional vector which contains the response functions with length N. Then the concurrent functional linear model can be defined in matrix notation as follows

$$\mathbf{y}(t) = \mathbf{Z}(t)\beta(t) + \epsilon(t). \tag{5.6}$$

Let the corresponding vector of residual functions with length N be defined as

$$\mathbf{r}(t) = \mathbf{y}(t) + \mathbf{Z}(t)\beta(t). \tag{5.7}$$

The following defines the weighted regularised fitted criterion

$$\text{LMSSE}(\beta) = \int \mathbf{r}(t)'\mathbf{r}(t)dt + \sum_{j}^{p} \lambda_j \int [L_j\beta_j(t)]^2 dt. \tag{5.8}$$

Let the following expansion

$$\beta_j(t) = \sum_{k}^{K_j} b_{kj}\theta_{kj}(t) = \theta_j(t)'\mathbf{b}_j$$

denote the regression function β_j regarding K_j basis functions denoted by θ_{kj}. Composite or supermatrices will have to be built so that Equations 5.6 and 5.8 can be conveyed in matrix notation referring clearly to these expansions (Ramsay *et. al* 2009).

Let $K_\beta = \sum_j^q K_j$, firstly vector $\mathbf{b}$ is built by arranging the vectors in a pile vertically with length K_β. This vector is given by $\mathbf{b} = (\mathbf{b}_1^{'}, \mathbf{b}_2^{'}, \ldots, \mathbf{b}_q^{'})'$. The matrix function $\Theta(t)$ is now put together with dimensions q by K_β and is given by

$$\Theta(t) = \begin{bmatrix} \theta_1(t)' & 0 & \cdots & 0 \\ 0 & \theta_2(t)' & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \theta_q(t)' \end{bmatrix}$$

Then $\beta(t) = \Theta(t)\mathbf{b}$ and now Equation 5.7 can be given as $\mathbf{r}(t) = \mathbf{y}(t) - \mathbf{Z}(t)\Theta(t)\mathbf{b}$. Now let $\mathbf{R}(\lambda)$ denote the block diagonal matrix with j^{th} block given by the following

$$\lambda_j \int [L_j\theta_j(t)]'[L_j\theta_j(t)]dt,$$

and then Equation 5.8 can be given as

$$\text{LMSSE}(\beta) = \int [\mathbf{y}(t)'\mathbf{y}(t) - 2\mathbf{b}'\Theta(t)'\mathbf{Z}(t)'\mathbf{y}(t) + \mathbf{b}'\Theta(t)'\mathbf{Z}(t)'\mathbf{Z}(t)\Theta(t)\mathbf{b}]dt + \mathbf{b}'\mathbf{R}(\lambda)\mathbf{b}. \tag{5.9}$$

To achieve the normal equations penalized least squares solution for the composite coefficient vector $\hat{\mathbf{b}}$ which are given as follows

$$[\int \Theta'(t)\mathbf{Z}'(t)\mathbf{Z}(t)\Theta(t)dt + \mathbf{R}(\lambda)]\hat{\mathbf{b}} = [\int \Theta'(t)\mathbf{Z}'(t)\mathbf{y}(t)dt], \tag{5.10}$$

Equation 5.9 is differentiated with respect to the coefficient vector denoted by $\mathbf{b}$ and the resulting derivative is set equal to zero. Equation 5.10 is a linear matrix equation determining the scalar coefficients in vector $\hat{\mathbf{b}}$, $\mathbf{A}\hat{\mathbf{b}} = \mathbf{d}$. Here the normal equation matrix denoted by $\mathbf{A}$ is given by

$$\mathbf{A} = \int \Theta'(t)\mathbf{Z}'(t)\mathbf{Z}(t)\Theta(t)dt + \mathbf{R}(\lambda),$$

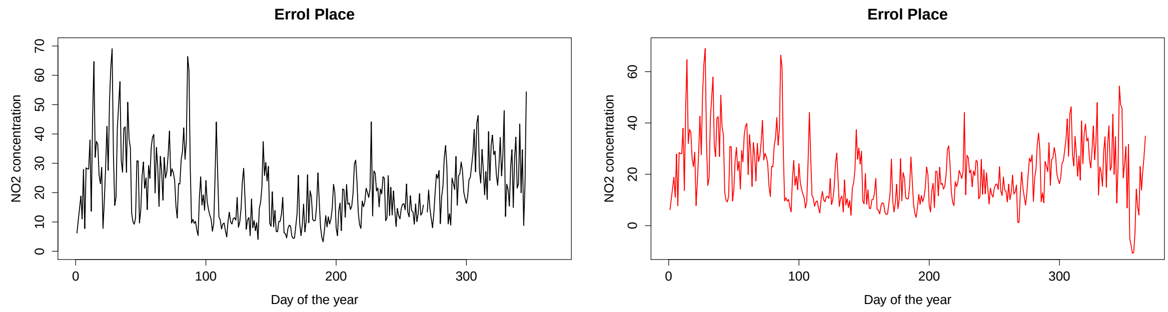
and $\mathbf{d} = \int \Theta'(t)\mathbf{Z}'(t)\mathbf{y}(t)dt$ (Ramsay *et. al* 2009).

In some cases the integrals can be clearly evaluated, although if not numerical integration will be returned too as this is both correct and practical in application (Ramsay *et. al*, 2009).

5.2.2 Missing Data

Monitoring Site Data

As mentioned previously in Chapter 2, missing data occurred in three of the monitoring sites out of six, namely Union Street Roadside, Errol Place and King Street. However, the percentage of missing data at these three sites were extremely low. Before producing the smoothed functions for the modelled and monitoring data, the missing values that occur in the monitoring data had to be dealt with. These missing values were dealt with using the *mnimput* function in R within the *mtsdi* package (CRAN, 2012). This function uses an altered version of the EM algorithm to impute missing values. When dealing with time series data, as in this case, missing values are evaluated considering both the correlation between time series and time structure of the series itself (CRAN, 2012). However, there were a few problems when imputing the missing values and this was mainly with Errol Place as it was at the end of the year values were missing, from day 347 to day 366. As it was at the end of the year the function was trying to impute missing values for, the results became unstable due to the lack of knowing the trend and pattern of the data after these missing values and also because the missing values were in a block. To highlight this a plot of Errol Place with missing data and a plot of Errol Place with imputations from the *mtsdi* package (CRAN, 2012) have been produced.



(a) Missing values included (b) Missing values imputed

Figure 5.1: Plots of Errol Place with both missing values included and missing values imputed

From Figure 5.1b, it is highlighted at the end of year that some of the predictions made at Errol Place are unstable and don't appear to fit well with the general trend throughout the year where some of the predictions made are negative. These negative predictions occur from day 354 to day 358 and these are not possible as an air quality measurement can not be negative. This issue was dealt with by simulating 5 values from the random normal distribution using the annual mean value of NO_2 concentrations at Errol Place as the mean and a standard deviation of 2.

Diffusion Tube Data

As mention in Chapter 2, missing data occurred in thirteen of the diffusion tubes out of forty. The percentage of missing data at these thirteen locations ranged from 8.33% to 25%. As this data was recorded monthly and in most cases there was only one missing value per diffusion tube, these values were imputed through using an average of the previous value and the value that occurred after. For example, if the data were recorded as 5, NA, 10 then the NA value would be recorded as 7.5.

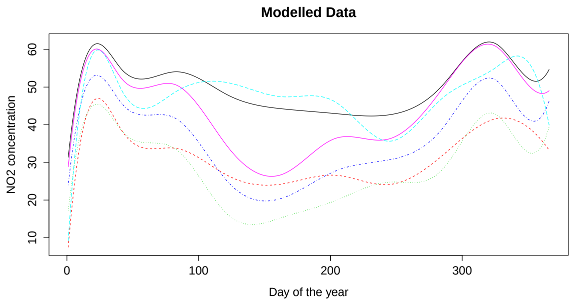
5.2.3 Selecting the Number of Basis Functions

A few issues emerged when it came to selecting the number of basis functions, the usual generalised cross validation (GCV) procedure that was outlined in Chapter 4 was no longer able to be used as it caused computation issues when running the model. Instead basis functions were chosen based on what aspects of the data were of interest and a few of these selections were looked at. This highlighted that slightly increasing and deceasing the amount of basis functions didn't change the overall results of the analysis. The values chosen here when modelling the monitoring and modelled data were 12, 24 and 36 where this highlights that 1, 2 and 3 basis functions were placed at each month of the year. However, when modelling the diffusion tube and modelled data, the values chosen were 4, 6 and 8.

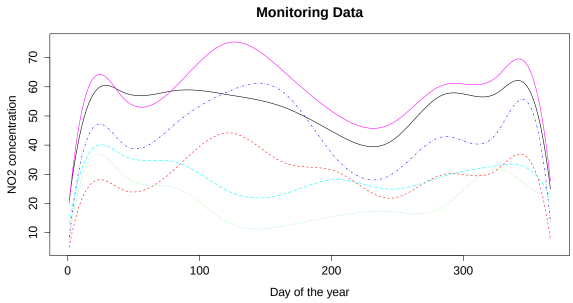
5.3 Results

5.3.1 Monitoring Data

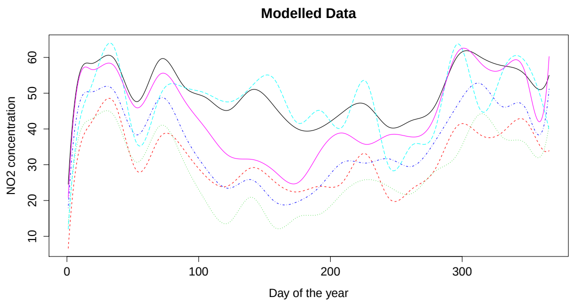
The main packages in *R* used to carry out this analysis were *fda* (CRAN, 2014) and *fda.usc* (CRAN, 2015b). When this analysis were carried out, daily NO_2 concentrations were used and data from the six monitoring site locations and the corresponding modelled data pixels closest to them were of interest. The number of basis functions used here are 12, 24 and 36 were 1, 2 and 3 basis functions have been placed respectively at each month of the year and the smoothed functions produced are as follows:



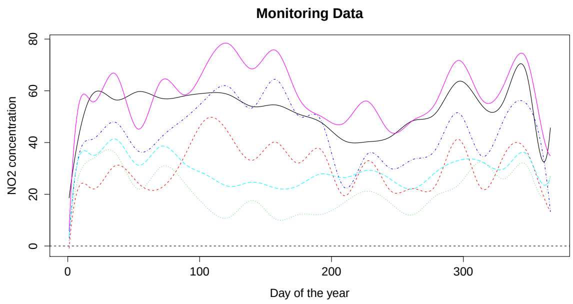
(a) Smoothed functions for the modelled data (12 basis functions)



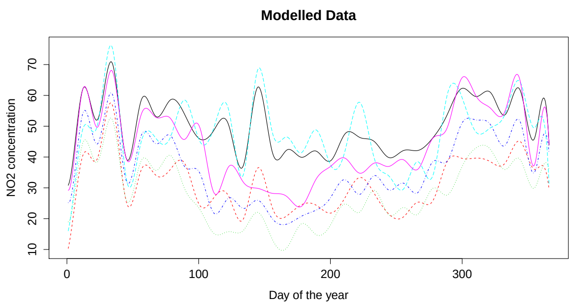
(b) Smoothed functions for the monitoring data (12 basis functions)



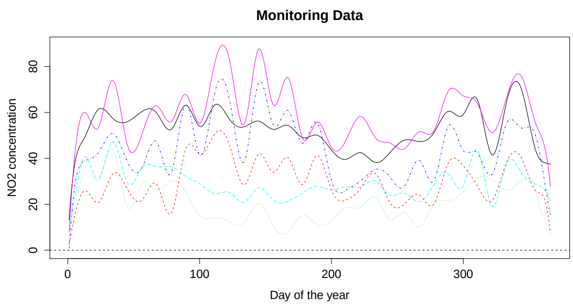
(c) Smoothed functions for the modelled data (24 basis functions)



(d) Smoothed functions for the monitoring data (24 basis functions)



(e) Smoothed functions for the modelled data (36 basis functions)

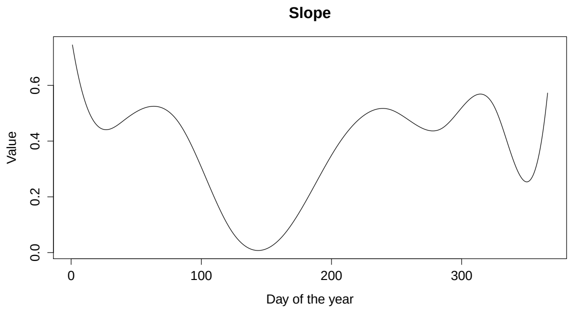


(f) Smoothed functions for the monitoring data (36 basis functions)

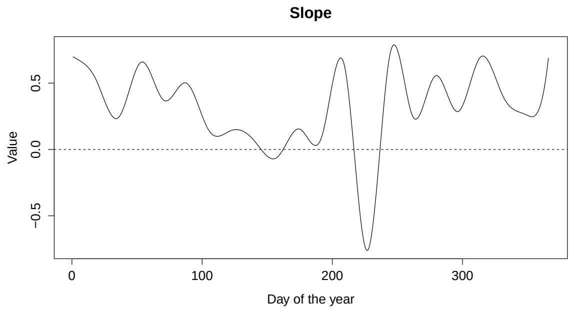
Figure 5.2: Smoothed functions for modelled and monitoring data where 12, 24 and 36 basis functions have been used

From Figure 5.2, it can be highlighted that the monitoring data could be said to be more variable as the range in the concentrations is slightly bigger as seen in Figures 5.2b, 5.2d and 5.2f were as for the modelled data the range in the concentrations is slightly smaller as seen in Figures 5.2a, 5.2c and 5.2e. The pattern in the functions day to day however, appear to be more variable in the modelled NO_2 data and the monitoring NO_2 data appears to be more constant from day to day.

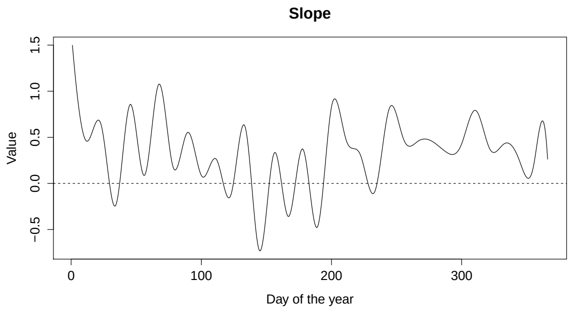
After running the concurrent functional linear model where the smoothed functions for the modelled data are the response variable and the smooth functions for the monitoring data are the explanatory variable, the following slope functions were produced:



(a) Slope function (12 basis functions)



(b) Slope function (24 basis functions)



(c) Slope function (36 basis functions)

Figure 5.3: Slope functions produced from running the concurrent functional linear model

Figure 5.3 highlights the slope functions produced from the functional regression model that has been carried out. From Figure 5.3a one can see that the slope varies from around 0 to 0.7 and at the peaks, the relationship between the modelled and monitoring data is thought to be stronger. From around day 100 (9^{th} of April) to day 200 (18^{th} of July) the value of the slope function is around zero highlighting that throughout this time period the modelled and monitoring data follow each other poorly. This can also be observed from Figure 5.3b, although it isn't as clear in Figure 5.3c. This reinforces what was highlighted in the time series plots in Chapter 2 of this thesis. It is suggested from these three slope functions that in general over the rest of the year the modelled and monitoring data appear to be well calibrated. In Chapter 2 when the monitoring site slopes were investigated individually through Deming Regression, the slopes ranged from 0.5 to 1.5 and for most of Figures 5.3b and 5.3c, this range can be observed. It was then of interest to add 95% confidence bands onto these three slopes as this would assist in seeing the uncertainty around them, the results are given below in Figures 5.4, 5.5 and 5.6.

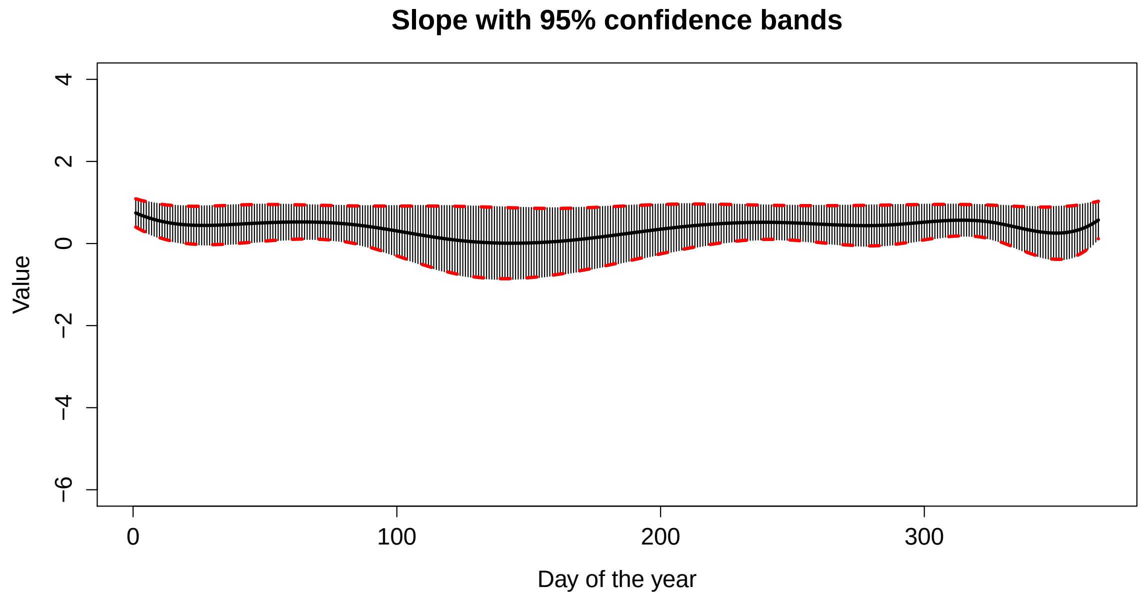


Figure 5.4: Slope function with 95% confidence bands (12 basis functions)

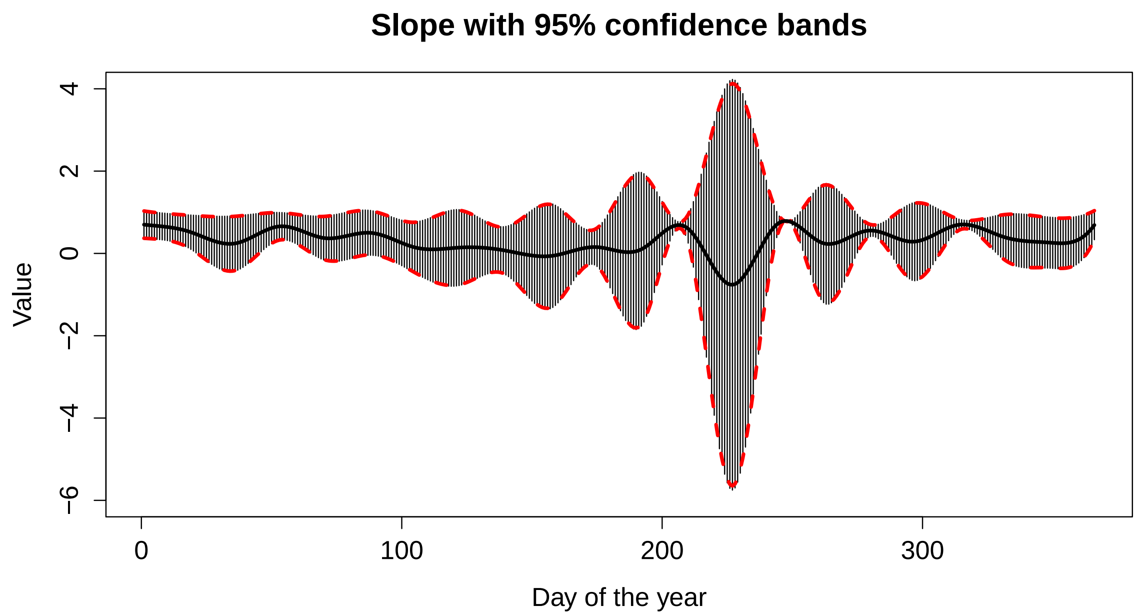


Figure 5.5: Slope function with 95% confidence bands (24 basis functions)

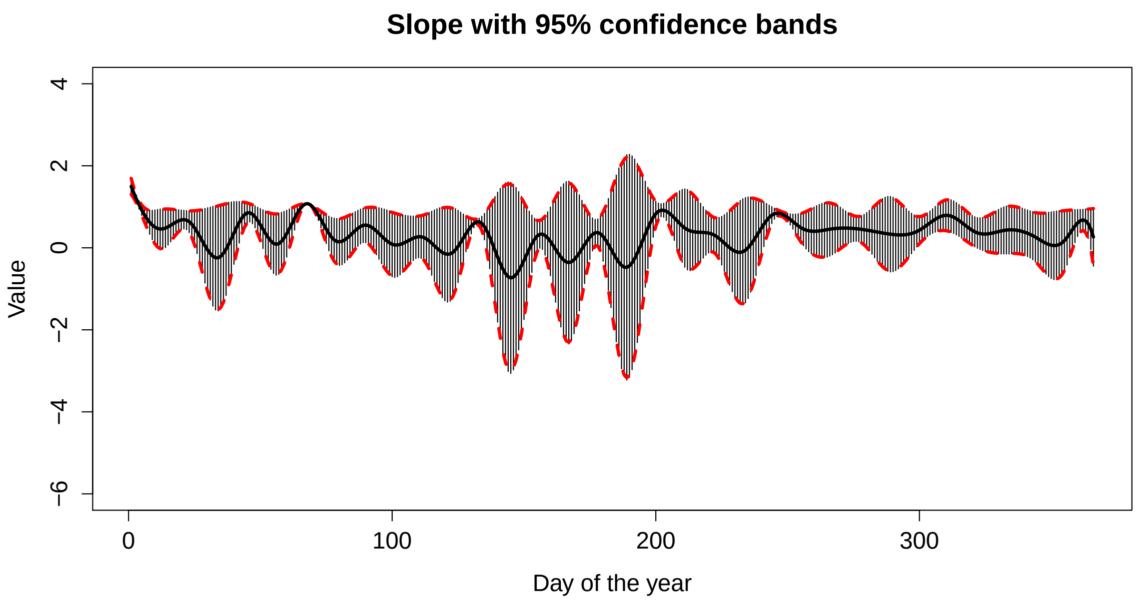
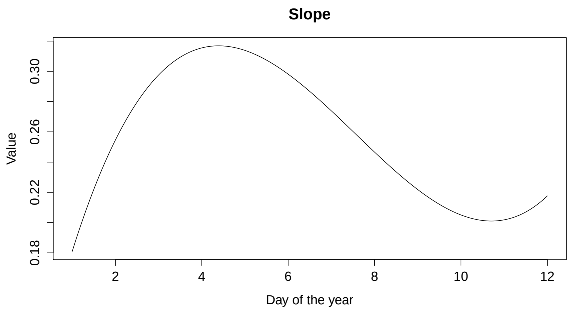


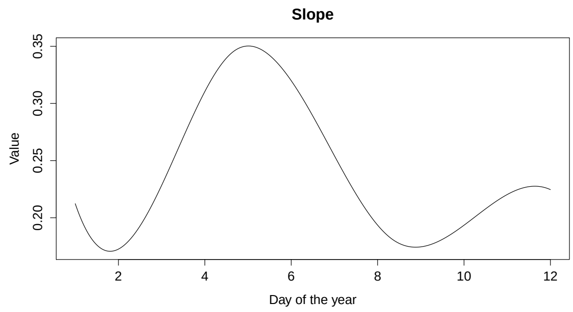
Figure 5.6: Slope function with 95% confidence bands (36 basis functions)

5.3.2 Diffusion Tube Data

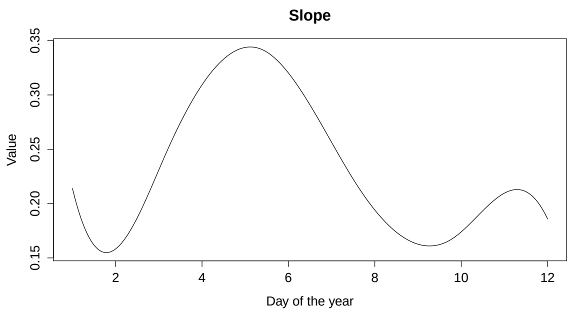
When the model was carried out for the diffusion tube data, monthly NO_2 concentrations were used and data from the forty diffusion tube locations and the corresponding modelled data pixels closest to them were of interest. The number of basis functions used here are 4, 6 and 8, once carrying out the concurrent functional linear model the following slope functions were produced:



(a) Slope function (4 basis functions)



(b) Slope function (6 basis functions)

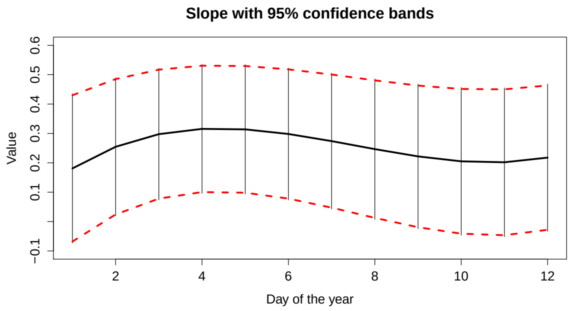


(c) Slope function (8 basis functions)

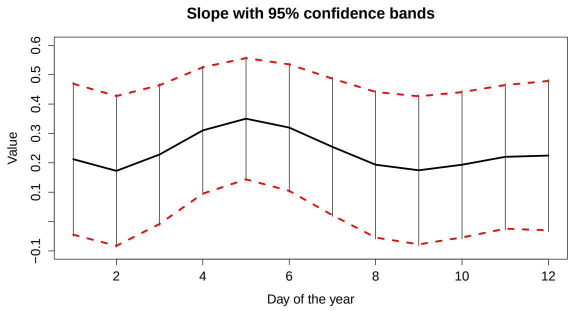
Figure 5.7: Slope functions produced from running the concurrent functional linear model

Figure 5.7 highlights the slope functions for different basis functions, produced from carrying out the functional regression model where the modelled data were the response variable and the diffusion tube data were the explanatory variable. It can be observed from Figure 5.7a that the slope varies from around 0.18 and 0.32 where as in Figure 5.7b the range of the slope varies from around 0.18 to 0.35 and this changes to around 0.15 to 0.35 in Figure 5.7c. This highlights that these ranges are all very similar and not changing much when the number of basis functions is changed. Again just like for the monitoring sites at the peaks, the relationship between the modelled and monitoring data is thought to be stronger. It can be highlighted in all three of these slope functions that the relationship between the diffusion tube data and the modelled data is strongest between the months of April to June. However, it should be observed that during these months the value of the slope is estimated to be between 0.32 and 0.35 depending on the number of basis functions and even though this is the peak value it still doesn't represent a very well calibrated relationship between both sets of data. When Deming Regression was carried out in Chapter 2, the overall slope value for the diffusion tube data was 0.3863 which is similar to what is being produced here and again emphasises that the modelled and diffusion tube data are not very well calibrated. Again, 95%

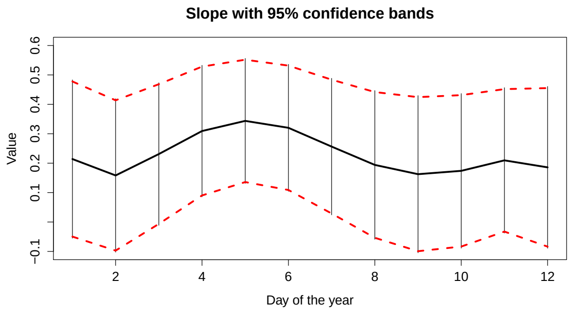
confidence bands were added onto these three slopes as this would assist in seeing the uncertainty around them, the results are given below in Figures 5.8a, 5.8b and 5.8c.



(a) Slope function with 95% confidence bands (4 basis functions)



(b) Slope function with 95% confidence bands (6 basis functions)



(c) Slope function with 95% confidence bands (8 basis functions)

Figure 5.8: Slope functions with 95% confidence bands produced from running the concurrent functional linear model

5.4 Conclusion

From this chapter, in conclusion it can be stated that through running functional regression it has been brought to the attention once again that the modelled and monitoring data are not very well calibrated from around the 9^{th} of April to 18^{th} of July. This has been highlighted throughout the analysis and this is thought to be due to temperatures of the sea being low at this time of the year. The atmosphere is stable as the air proceeds over the cold sea, however the air begins to become unstable by spring time as the sun is becoming more powerful and warms up the ground. This is a procedure known as fumigation and this is not an option given for the ADMS-Urban model set up and can not be taken into consideration whilst running the model. It's plausible that given Aberdeen is situated on the coast, and if there is an onshore breeze in spring, the air is fairly low in temperature as it comes off the North sea. Therefore mixing is not as high as would be expected for an inland location, hence measured pollutant concentrations may be higher than predicted by the model.

From the slope functions it was also seen that the overall monitoring and modelled data appear to be reasonably well calibrated throughout the rest of the year 2012 with no other main features standing out. Also investigating the slope functions for the modelled data against the diffusion tube data it is suggested that both of these sets of data are well calibrated between the months of April and June. Even though between those months the value of the slope function is at its highest, the value isn't very high (0.35) suggesting that overall the diffusion tube data and modelled data do not appear to follow each other very well.

Chapter 6

Conclusion and Discussion

6.1 Introduction

The main aim of this research was to examine various statistical methods to compare the output of the ADMS-Urban model with monitoring site and diffusion tube NO_2 data over the region of Aberdeen. Aberdeen has six automatic monitoring sites within the city, these are located in Wellington Road, Market Street, Anderson Drive, Union Street Roadside, Errol Place and King Street and there are 46 diffusion tubes located in Aberdeen. However, for this thesis data for 40 of these locations were available. As previously stated air quality modelling is a crucial tool for expanding and assessing air quality policy (DEFRA, 2011a). The ADMS-Urban model run used throughout this thesis was set up with 181 road sources and no industrial or grid sources for 2012. The chemistry module was turned on and buildings and complex terrain was turned off. Details of the road were incorporated, these included width of the road, canyon height and elevation of the road. Time varying emission factors were also incorporated into the model run, for example this allowed for weekdays to be different from Saturday and Sunday. Monitoring data from Errol Place was used for the background concentrations suggesting that the modelled and monitoring data at this site will appear to be calibrated better. The ADMS-Urban modelled data was of high resolution meaning that a very accurate representation of the air quality concentrations in Aberdeen were possible. The background pixels of the ADMS-Urban model were 75 m $\times$ 75 m apart and also included pixels for the roads in Aberdeen, although the roads were slightly different as the grid spacing were not equidistant. In total there were 18319 pixels and 10201 of these pixels were background pixels. For each one of these pixels hourly data were recorded over 2012, this gave 8784 observations for each pixel.

The main aim of this research was split up into three sub aims, the first of these was to investigate for each monitoring site point and overall diffusion tube data, how comparable the ADMS-Urban model and observed data were. The second aim was to explore how comparable the ADMS-Urban modelled and observed data are over the full domain of Aberdeen. The final aim was to examine FDA techniques to analysis the characteristics of the ADMS-Urban modelled pixels in space and also to see in a functional context how comparable the ADMS-Urban modelled data and the observed data are.

6.2 Monitoring site and model comparison

The first of these aims was achieved through the use of techniques such as plotting the differences between the modelled and monitoring site data, Deming Regression, Bland Altman plots and extreme value analysis. These techniques were used to explore the effect of uncertainties in both the modelled and monitored data, and to also examine the similarity in patterns of exceedances. From these investigations it was found that the ADMS-Urban modelled data and monitoring data at Wellington Road are not very well calibrated over the year 2012 compared with the other monitoring sites in Aberdeen. This was suggested through the difference plots where Wellington Road had the highest differences occurring with some as large as 100 μgm^{-3} and also through Deming Regression which highlighted for every 1 μgm^{-3} increase in the monitoring data, on average the modelled data increases by 0.4998 μgm^{-3}. Through the use of bland altman plots it was highlighted that as the variation in differences increased, the average also increased. This was seen at Wellington Road, Anderson Drive and Market Street, where as at Errol Place and King Street there appeared to more of a linear relationship between the average against differences. Although the model appeared to perform poorly at Wellington Road, at the other monitoring stations the modelled and monitoring data appeared to be roughly well calibrated.

Comparing the diffusion tube and modelled data it appeared that overall these data were not well calibrated over 2012 with deming regression highlighting that for every 1 μgm^{-3} increase in the diffusion tube NO_2 data, on average the modelled data increases by 0.3863 μgm^{-3}. However, the diffusion tube data are collected irregularly though approximately monthly. This means when we are comparing these data to the modelled data, the values are slightly mismatched temporally. This may be the reason why both the modelled and diffusion tube data are not very well calibrated when exploring the different techniques. Exploring the number of exceedances over the 90^{th} percentile for both the modelled and monitoring data suggested that compared with the summer, in the winter the monitoring sites appear to have a similar amount of exceedances for both the modelled and monitoring data. The number of exceedances over the 75^{th}, 95^{th} and 99^{th} percentile were also investigated, however the 90^{th} percentile was chosen to explore further. This percentile was chosen as there appeared to be a sufficient number of exceedances to observe if the modelled and monitoring data were occurring at the same points in time over 2012. Different thresholds were determined through mean residual life plots for the modelled and monitoring data. Determining the number of events that exceed these thresholds, it can be concluded that more events exceeded in the monitoring data in most cases highlighting there is more variability in the monitoring data as they are exceeding the given thresholds more. This suggests that the monitoring data are more likely to observe and pick up on larger NO_2 concentrations than the modelled data.

6.3 Spatial Comparison

The second aim was investigated through the use of statistical spatial models, kriging and investigating the differences between the observed and modelled data. Throughout this analysis the main challenge faced was the size of the ADMS-Urban modelled data which led to computational challenges. To help to resolve this challenge when carrying out these analyses, the main city centre region of Aberdeen was focussed on to reduce the size of the data. This area were chosen as this is where most of the road traffic is in Aberdeen and road traffic is thought to be the main cause of concern in terms of air quality in Aberdeen. As previously mentioned it is stated by the Aberdeen City Council (2013) that atmospheric pollution in the city is mainly caused by road traffic (Aberdeen City Council, 2013). These approaches highlighted that the model appears to perform relatively well over the region of Aberdeen with the model emphasising that the area in Aberdeen with the highest NO_2 concentrations is towards the east. This is hardly surprising as this is where the city centre and Harbour are situated. The annual modelled predictions suggested that NO_2 concentrations are higher on the roads of

Aberdeen and in fact the roads appeared to dominate the pollutant predictions. This was further shown in the background pixels of the annual modelled predictions as they still highlighted the spatial structure of the roads.

As well as annual predictions, monthly predictions were also investigated. The monthly modelled, diffusion tube and monitoring site spatial models highlighted that the NO_2 predictions do not vary over the months. A reason for this could have been due to the lack of data in the observed data case as we only have data for six monitoring sites and thirty eight diffusion tubes locations. This leads to limited spatial coverage over the entire region of Aberdeen, all monitoring sites and most of the diffusion tubes are located around the main city centre of Aberdeen (East of Aberdeen) and in the West of the city, air quality doesn't appear to be monitored. It would have been beneficial to have at least one monitoring site/diffusion tube situated in the West of Aberdeen in order to compare the modelled and observed data outwith the city centre. Even though the monthly predictions didn't appear to change much over the year for both the modelled and observed data there appeared to be a region in space that was always predicted to have higher NO_2 concentrations. Similar to the annual predictions this region was in the city centre of Aberdeen. Investigating the annual predicted differences between the modelled and diffusion tube/monitoring site surfaces, it was concluded that the differences were rather small ranging from -0.7 μgm^{-3} to -0.1 μgm^{-3}. These differences were explored for the modelled data including and excluding the pixels that represent the roads and in both cases these differences were small. These differences also suggested spatial structure within the roads and the variograms in both cases highlighted strong residual spatial correlation.

6.4 Dimension reduction and common behaviours in ADMS-Urban model

The final aim was achieved through the use of FPCA, clustering and functional regression. By carrying out this analysis we would be able to determine how the pixels behaved in space and if there was any patterns occurring in space. We would also be able to determine an overall temporally varying slope for the relationship between modelled and observed data and this would highlight overall how comparable these data are. Clustering emphasised the dominance of the roads in Aberdeen as again the modelled data examined included and excluded the pixels that represented the roads. Even once the pixels that represented the roads had been removed there still appeared to be spatial structure indicating the roads. This emphasises that in Aberdeen in terms of air quality, road traffic is a major concern. Cluster mean curves highlighted that Wellington Road didn't perform as well as the other monitoring sites, as the daily mean monitoring data was much higher than the cluster mean curves. Examining the cluster mean curves for the winter and summer months emphasised that in the winter the monitoring site data followed the cluster mean curves more closely.

Functional regression indicated that the modelled and monitoring data are not very well calibrated from around mid April to around mid July as the value of the slope function was around zero. This was also suggested in Chapter 2 when plotting the time series plots of the modelled and monitoring data and comparing them. The reason for this is thought to be driven by temperatures of the sea being low at this time of the year. There is a process known as fumigation, which is not an option in the model set up, and hence is not taken into consideration whilst running the model. This process is when the atmosphere is stable as the air proceeds over the cold sea, however the air begins to become unstable by spring time as the sun is becoming more powerful and warms up the ground. It is very likely that as Aberdeen is located on the coast, and if there is an onshore breeze in spring, the air is fairly low in temperature as it comes off the North sea. Hence, mixing is not as high as would be expected for an inland location leading to measured pollutant concentrations potentially being higher than predicted by the model. Overall the slope function ranged from around -0.5 to 1.5 with no other main features appearing to stand out of the slope functions and it was concluded that over for the rest of 2012 the monitoring and modelled data appeared to be relatively well calibrated. Overall the slope functions produced from modelling the diffusion tube and modelled data highlighted that over 2012 these data do not appear to be very comparable with a slope value of 0.35 at its highest.

6.5 Further Work

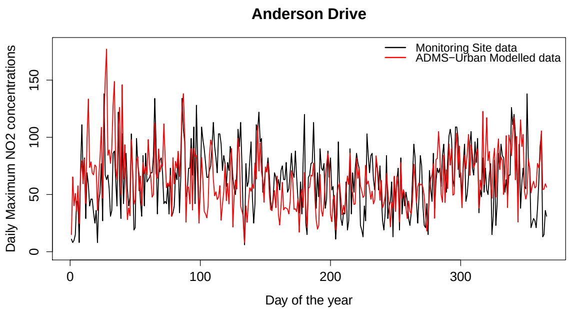
For this thesis, investigations were made based only on the year 2012, given more time it would be beneficial to compare this with other years or to consider more years. This will highlight whether the modelled and observed data are comparable over other years and how they both change from year to year and if these differences and changes are similar. Throughout this thesis the pollutant NO_2 was focussed on, ADMS-Urban modelled data and monitoring data are also available for NO_x, PM_{10} and $PM_{2.5}$. It would be of interest to use the various techniques to observe how well the modelled and monitoring data are calibrated for the other pollutants. This would highlight if the modelled and monitoring data are more comparable for certain pollutants.

Since there is uncertainty in both measurement and model a sensitivity analysis for the model could be useful. This would give insight into how robust the results are when the inputs of the model are changed. We could also investigate handling correlations in both space and time and thus building a spatio-temporal model for NO_2 concentrations. Finally, it would also be of interest to consider an air quality indicator, since measures are put in place to control and manage air quality. Both the ADMS-Urban model and the monitoring data could be used to firstly test through simulation, the effect of proposed changes and secondly provide an assessment of the effect, if any, of the management measures.

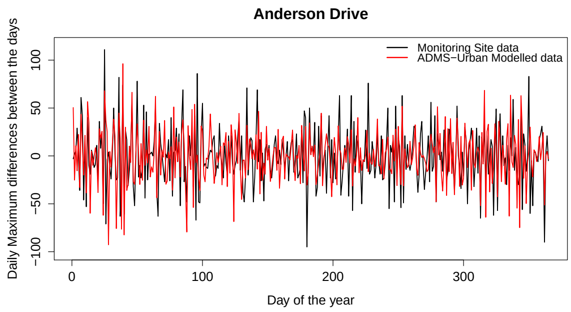
Appendices

Appendix A

Daily Maximum and Daily Maximum Difference between the days plots



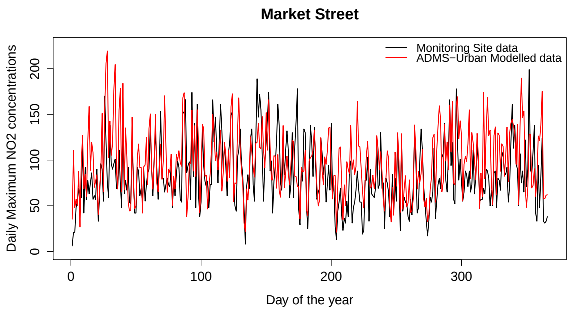
(a) Daily Maximum NO_2 data



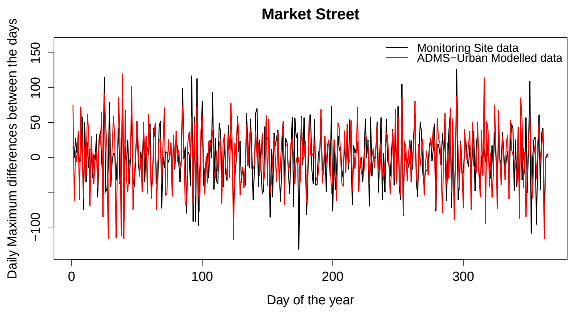
(b) Daily Maximum difference between the days

Figure A.1: Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Anderson Drive (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

From Figure A.1a, the daily maximum profile plot highlights that at Anderson Drive, the modelled data appears to be higher than the monitoring data at the very start of the year. Then conversely it seems to be lower than the monitoring data throughout most of the year. This is until the latter part of the year where it seems to be slightly higher than the monitoring data on certain days. Both the daily maximum differences day to day in the monitoring and modelled data appear to be varied over the short timescale. This is highlighted through the large differences shown in Figure A.1b.



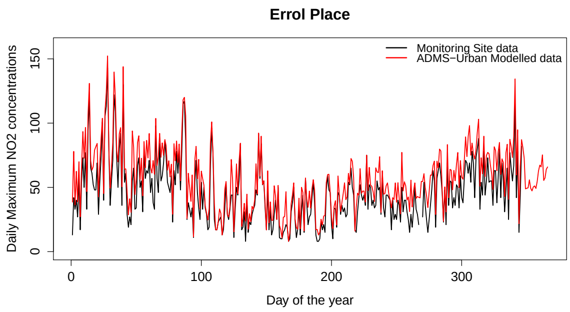
(a) Daily Maximum NO_2 data



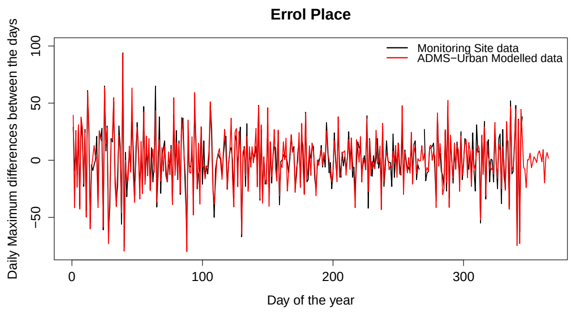
(b) Daily Maximum difference between the days

Figure A.2: Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Market Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

Figure A.2a highlights that at Market Street the modelled data appears to be higher than the daily maximum NO_2 monitoring concentrations at the beginning and end of 2012. The modelled data also appears to have more variability than the monitoring data and this appears more obvious from Figure A.2b as the day to day differences appear to be bigger.



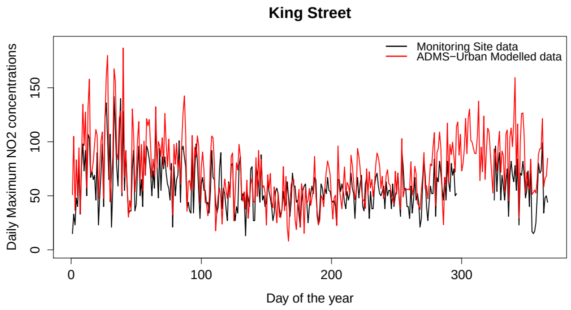
(a) Daily Maximum NO_2 data



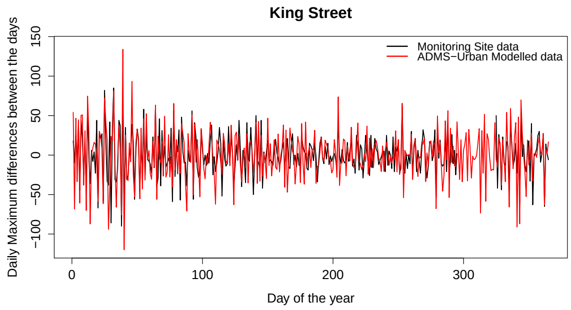
(b) Daily Maximum difference between the days

Figure A.3: Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site Errol Place (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

As previously, Figure A.3a highlights that at Errol Place the modelled data appears to be slightly higher than the monitoring NO_2 concentrations throughout 2012. The daily maximum NO_2 modelled data also seem much more varied over 2012. This can be suggested due to the daily maximum difference between days plot of NO_2 concentrations in Figure A.3b as the differences are bigger for the modelled data.



(a) Daily Maximum NO_2 data



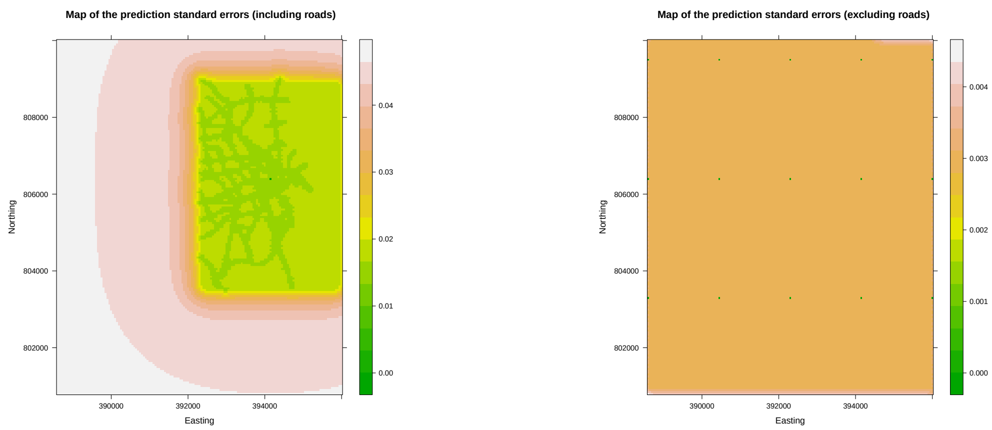
(b) Daily Maximum difference between the days

Figure A.4: Daily Maximum and Daily Maximum Difference between Days Time Series Plots of NO_2 concentration at the monitoring site King Street (Monitoring site data is represented by the black line and the ADMS-Urban modelled data is represented by the red line)

The same pattern occurs here for the daily maximum NO_2 concentrations that occurred for the daily mean NO_2 concentrations at King Street. Figure A.4b highlights that the modelled data appear to have more variability than the monitoring observations and this more obvious at the start of the year as the day to day differences are notably bigger. Figure A.4a highlights that the modelled data appears to be higher than the monitoring data at King Street.

Appendix B

Map of the annual modelled prediction standard errors

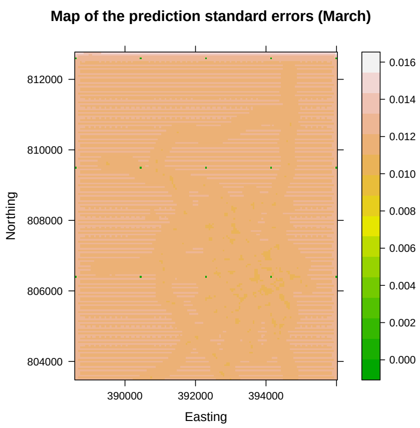


(a) Including roads (b) Excluding roads

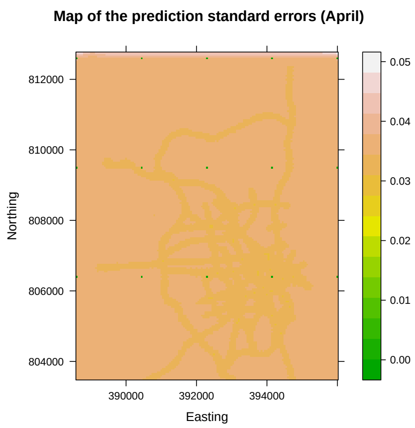
Figure B.1: Map of the annual modelled standard errors (Including and Excluding the roads)

Appendix C

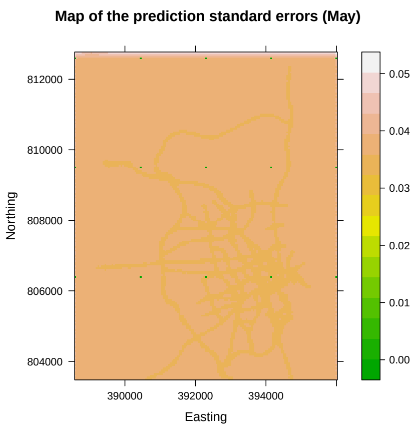
Map of the monthly modelled prediction standard errors



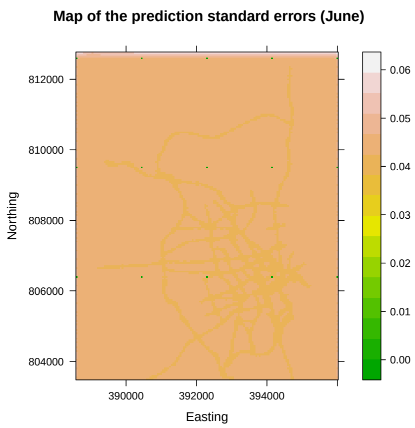
(a)



(b)



(c)



(d)

Figure C.1: Map of the modelled prediction standard errors (March to June)

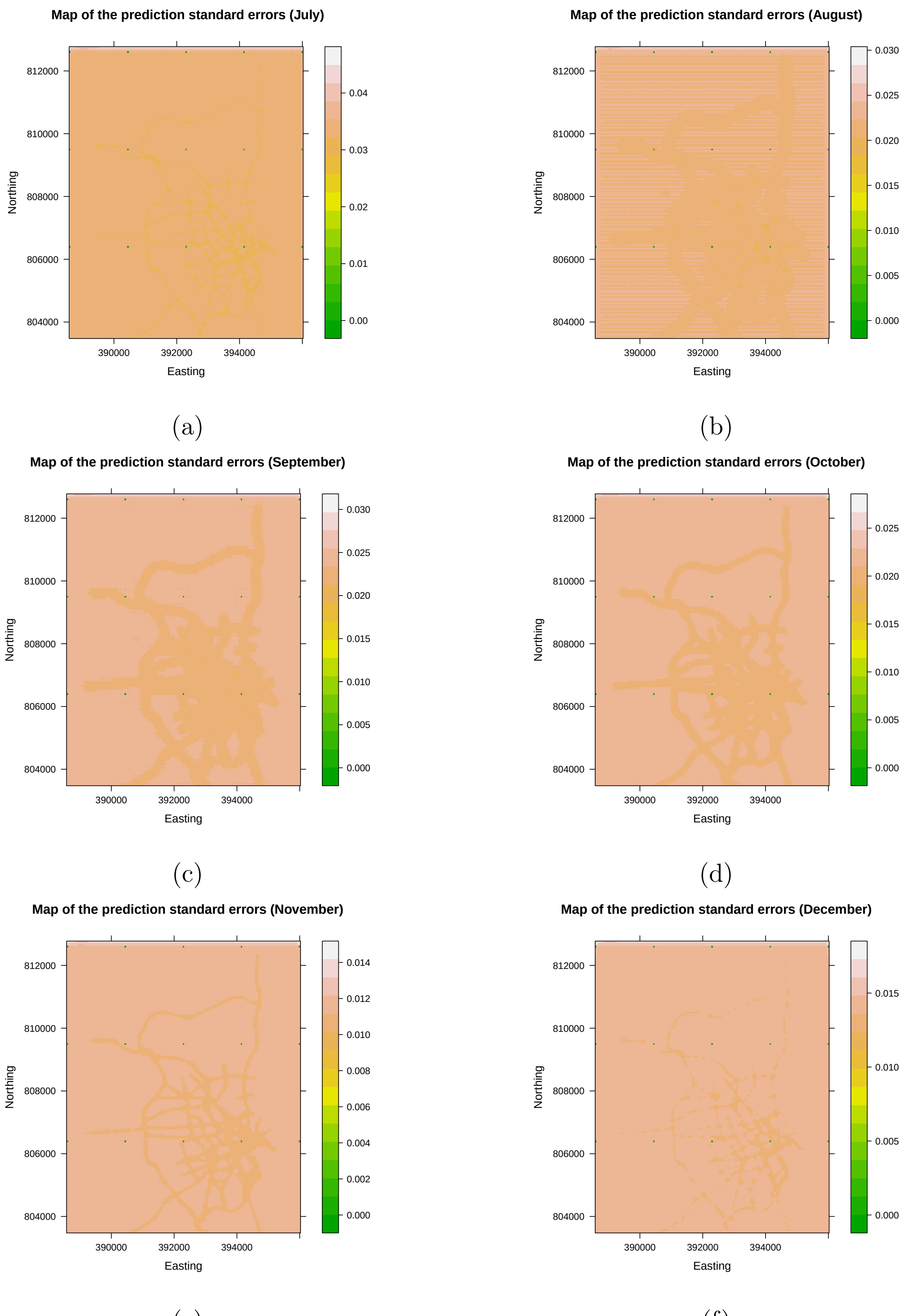


Figure C.2: Map of the modelled prediction standard errors (July to December)

Bibliography

Aberdeen City Council (2013). Air Quality Progress Report for Aberdeen City Council. Available at: http://www.aberdeencity.gov.uk/nmsruntime/saveasdialog.asp?lID=53090&sID=5034 [Accessed 15/02/15].

Abraham, C., Cornillon, P.A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 30*, 581-595.

Ahmadi, S.H. and Sedghamiz, A. (2007). Geostatistical Analysis of Spatial and Temporal Variations of Groundwater Level. *Environmental Monitoring and Assessment 129*, 277-294.

Air Quality in Scotland (2014a). Air Quality Standards and Objectives. Available at: http://www.scottishairquality.co.uk/air-quality/standards [Accessed 23/10/14].

Air Quality in Scotland (2014b). Data Selector. Available at: http://www.scottishairquality.co.uk/data/data-selector [Accessed 6/10/14].

Air Quality in Scotland (2014c). Monitoring site locations. Available at: http://www.scottishairquality.co.uk/air-quality/monitoring [Accessed 23/10/14].

Air Quality in Scotland (2015). Latest pollution map. Available at: http://www.scottishairquality.co.uk/latest/?site_id=ABD&view=latest [Accessed 13/02/15].

Air Quality Modelling Review Steering Group. (2011). Review of Air Quality Modelling in DEFRA. Available at: http://uk-air.defra.gov.uk/assets/documents/reports/cat20/1106290858_DEFRAModellingReviewFinalReport.pdf [Accessed 27/03/15].

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., and Andreassian, V. (2012). Characterising performance of environmental models. *Environmental Modelling & Software 40*, 1-20.

Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. (*Lancet*, 1986; **i:** 307-310), 1-9.

CEH (2014). Centre for Ecology and Hydrology. Available at: http://www.ceh.ac.uk/ [Accessed 24/10/14].

CERC (2013). ADMS-Urban, Urban Air Quality Management System. *Version 3.2*, 243-257.

CERC (2014a). Air pollution modelling. Available at: http://www.cerc.co.uk/environmental-software.html [Accessed 23/10/13].

CERC (2014b). ADMS 5. Available at: http://www.cerc.co.uk/environmental-software/ADMS-model.html [Accessed 23/10/14].

CERC (2014c). ADMS-Urban. Available at: http://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html [Accessed 24/10/14].

CERC (2014d). ADMS-Roads (Extra). Available at: http://www.cerc.co.uk/environmental-software/ADMS-Roads-model.html [Accessed 24/10/14].

CERC (2014e). ADMS-Airport. Available at: http://www.cerc.co.uk/environmental-software/ADMS-Airport-model.html [Accessed 24/10/14].

CERC (2014f). ADMS-Screen. Available at: http://www.cerc.co.uk/environmental-software/ADMS-Screen-model.html [Accessed 24/10/14].

CERC (2014g). ADMS-Urban, Model input data and Model output data. Available at: http://www.cerc.co.uk/environmental-software/ADMS-Urban-model/data.html [Accessed 10/11/14].

Charrad M., Ghazzali N., Boiteau V., and Niknafs A. (2014). NbClust: An R Package for Determing the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software 61*, 1-36.

Clean Air Act (July, 1956). Clean Air Act, 1956. Available at: http://www.legislation.gov.uk/ukpga/1956/52/pdfs/ukpga_19560052_en.pdf [Accessed 14/01/15].

Coles, S., and Davison, A. (2008). Statistical Modelling of Extreme Values. Available at: http://www.cces.ethz.ch/projects/hazri/EXTREMES/talks/colesDavisonDavosJan08.pdf [Accessed 27/03/15].

COMEAP (June, 2011). Review of the UK air quality index, pp 9-11. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/304633/COMEAP_review_of_the_uk_air_quality_index.pdf [Accessed 23/10/14].

CRAN (2012). Package "mtsdi". Multivariate time series data imputation. Available at: https://cran.r-project.org/web/packages/mtsdi/mtsdi.pdf [Accessed 30/07/15].

CRAN (2013a). Package "MethComp". Functions for Analysis of Agreement in Method Comparison Studies. Available at: https://cran.r-project.org/web/packages/MethComp/MethComp.pdf [Accessed 27/02/16].

CRAN (2013b). Package "fExtremes". Rmetrics - Extreme Financial Market Data. Available at: https://cran.r-project.org/web/packages/fExtremes/fExtremes.pdf [Accessed 27/02/16].

CRAN (2014). Package "fda". Functional Data Analysis. Available at: https://cran.r-project.org/web/packages/fda/fda.pdf [Accessed 27/02/16].

CRAN (2015a). Package "geoR". Analysis of Geostatistical Data. Available at: https://cran.r-project.org/web/packages/geoR/geoR.pdf [Accessed 27/02/16].

CRAN (2015b). Package "fda.usc". Functional Data Analysis and Utilities for Statistical Computing. Available at: https://cran.r-project.org/web/packages/fda.usc/fda.usc.pdf [Accessed 27/02/16].

CRAN (2015c). Package "fpc". Flexible Procedures for Clustering. Available at: https://cran.r-project.org/web/packages/fpc/fpc.pdf [Accessed 27/02/16].

CRAN (2015d). Package "cluster". "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. Available at: https://cran.r-project.org/web/packages/cluster/cluster.pdf [Accessed 27/02/16].

CRAN (2015e). Package "ggplot2". An Implementation of the Grammar of Graphics. Available at: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf [Accessed 27/02/16].

CRAN (2016a). Package "gstat". Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation. Available at: https://cran.r-project.org/web/packages/gstat/gstat.pdf [Accessed 27/02/16].

CRAN (2016b). Package "sp". Classes and Methods for Spatial Data. Available at: https://cran.r-project.org/web/packages/sp/sp.pdf [Accessed 27/02/16].

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377-403.

Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *The Canadian Journal of Statistics 30*, 285-300.

de Boor, C. (2001). A Practical Guide to Splines. Revised Edition. New York: Springer.

DEFRA (2007). The Air Quality Strategy for England, Scotland, Wales and Northern Ireland. *Volume 1*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/69336/pb12654-air-quality-strategy-vol1-070712.pdf [Accessed 27/10/14].

DEFRA (2008). Diffusion Tubes for Ambient NO_2 Monitoring: Practical Guidance for Laboratories and Users. *Issue 1a*. Available at: http://laqm.defra.gov.uk/documents/0802141004_NO2_WG_PracticalGuidance_Issue1a.pdf [Accessed 24/02/15].

DEFRA (May, 2009). Frame. Available at: http://pollutantdeposition.defra.gov.uk/frame [Accessed 24/10/14].

DEFRA (2011a). Air Quality Modelling. Available at: http://uk-air.defra.gov.uk/research/air-quality-modelling [Accessed 24/10/14].

DEFRA (2011b). UK and EU Air Quality Policy Context. Available at: http://uk-air.defra.gov.uk/air-pollution/uk-eu-policy-context [Accessed 25/10/14].

DEFRA (2012). Automatic Urban and Rural Network (AURN). Available at: http://uk-air.defra.gov.uk/networks/network-info?view=aurn [Accessed 24/02/15].

DEFRA (March, 2013). Air Modelling for DEFRA. Available at: http://uk-air.defra.gov.uk/research/air-quality-modelling?view=modelling [Accessed 24/10/14].

DEFRA (2014a). Forecast maps (provided by the Met Office). Available at: http://uk-air.defra.gov.uk/forecasting/ [Accessed 24/10/14].

DEFRA (2014b). Background Mapping data for local authorities. Available at: http://uk-air.defra.gov.uk/data/laqm-background-home. [Accessed 16/11/14].

DEFRA (February, 2014). Modelled air quality data. Available at: http://uk-air.defra.gov.uk/data/modelling-data [Accessed 24/10/14].

DEFRA (March, 2014). Protecting and enhancing our urban and natural environment to improve public health and wellbeing. Available at: https://www.gov.uk/government/policies/protecting-and-enhancing-our-urban-and-natural-environment-to-improve-public-health-and-wellbeing [Accessed 23/10/14].

Deserti, M., Cacciamani, C., Golinelli, M., Kerschbaumer, A., Leoncini, G., Savoia, E., Selvini, A., Paccagnella, T., Tibaldi, S. (2001). Operational meteorological pre-processing at Emilia-Romagna ARPA Meteorological Service as a part of a decision support system for air quality management. In: Coppalle, A. (Ed.), Proceedings of the Sixth Workshop on Harmonisation Within Atmospheric Dispersion Modelling for Regulatory Purposes. *International Journal of Environment and Pollution 16 (1-6)*, 571-582.

Diggle, P.J., and Ribeiro Jr., P.J. (2007). Model-based Geostatistics. New York: Springer Science+Business Media, LLC.

Environmental Audit Committee (2011). Air quality: A follow up report - Environmental Audit Committee. Available at: http://www.publications.parliament.uk/pa/cm201012/cmselect/cmenvaud/1024/102403.htm [Accessed 14/1/15].

Esri. (2015). What is GIS? Available at: http://www.esri.com/what-is-gis [Accessed 27/03/15].

Eubank, R. L. (1999). Spline Smoothing and Nonparametric Regression, Second Edition. New York: Marcel Dekker.

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society. Series B (Methodological) 65*, 57-80.

Faraway, J.J. (1997). Regression Analysis for a Functional Response. *American Statistical Association and the American Society for Quality Control TECHNOMETRICS 39*, 254-261.

Friedman, J. H., and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Journal of the American Statistical Association 31*, 1-39.

Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *Revstat 8*, 57-80. Available at: http://www.ine.pt/revstat/pdf/rs100104.pdf [Accessed 10/03/15].

Green, P. J., and Silverman, B. W. (1994). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. London: Chapman and Hall.

Hartigan, J.A., and Wong, M.A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society C 28*, 100-108.

Hastie, T., and Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological) 55*, 757-796.

Heimann, I., Bright, V.B., McLeod, M.W., Mead, M.I., Popoola, O.A.M., Stewart, G.B., and Jones, R.L. (2015). Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. *Atmospheric Environment 113*, 10-19.

Holtslag, A.A.M., Van Ulden, A.P. (1983). A simple scheme for daytime estimates of the surface fluxes from routine weather data. *Journal of Climate and Applied Meteorology 22*, 517-529.

Ignaccolo, R., Ghigo, S., and Giovenali, E. (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics 19*, 672-686.

Kaufman, L., and Rousseeuw, P.J. (1990). Finding groups in data: An Introduction to Cluster Analysis. John Wiley & Sons, New York, 1990.

Linnet, K. (1990). Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statistics in Medicine 9*, 1463-1473.

Mabbett (2014). Air Dispersion Modelling. Available at: http://www.mabbett.eu/services/environment/air_dispersion_modelling [Accessed 28/10/14].

MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In LML Cam, J Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, 281-297.

Matejicek, L. (2014). Using Geostatistical Tools for Mapping Traffic-Related Air Pollution in Urban Areas. *International Environmental Modelling and Software Society 2*, 1031-1036.

National Atmospheric Emissions Inventory. (2014). UK NAEI - National Atmospheric Emissions Inventory. Available at: http://naei.defra.gov.uk/ [Accessed 27/03/15].

Oberkampf, W.L., and Barone, M.F. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics 217*, 5-36.

Pannullo, F., Lee, D., Waclawski, E., and Leyland, A.H. (2015). Improving spatial nitrogen dioxide prediction using diffusion tubes: A case study in West Central Scotland. *Atmospheric Environment 118*, 227-235.

Ramsay, J.O., Hooker, G., and Graves, S. (2009). Functional Data Analysis with R and MATLAB. New York: Springer Science+Business Media, LLC.

Ramsay, J.O., and Silverman, B.W. (2005). Functional Data Analysis. New York: Springer Science+Business Media, Inc.

Ricardo-AEA (2013). Ricardo-AEA. Available at: http://www.ricardo-aea.com/cms/ [Accessed 24/10/14].

Righi, S., Lucialli, P., and Pollini, E. (2009). Statistical and diagnostic evaluation of the ADMS-Urban model compared with an urban air quality monitoring network. *Atmospheric Environment 43*, 3850-3857.

Scire, J.S., Robe, F.R., Fernau, M.E., Yamartino, R.J. (2000). A User Guide for the CALMET Meteorological Model (Version 5). Earth Tech, Inc.

Sokhi, R.S., San Josè, R., Kitwiroon, N., Fragkou, E., Pérez, J.L., and Middleton, D.R. (2005). Prediction of ozone levels in London using MM5-CMAQ modelling system. *Environmental Modelling & Software 21*, 566-576.

Stockholm (2014). Uppsala County Air Quality Management Association monitors regions air quality. Available at: http://www.slb.nu/elvf/ [Accessed 23/10/14].

van der Laan, M.J., Pollard, K.S., and Bryan J. (2002). A New Partitioning Around Medoids Algorithm. Available at: http://biostats.bepress.com/cgi/viewcontent.cgi?article=1003&context=ucbbiostat [Accessed 15/06/15].

Wilmott, C.J. (1982). Some comments on the evaluation of model performance. *Bulletin American Meteorological Society 63*, 1309-1313.

Wong, D.W., Yuan, L., and Perlin, S.A. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Analysis and Environmental Epidemiology 14*, 404-415.

Yen, J.D.L., Thomson, J.R., Paganin, D.M., Keith, J.M., and Mac Nally, R. (2015). Function regression in ecology and evolution: FREE. *Methods in Ecology and Evolution 6*, 17-26.